Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deciphering Wood Pyrolysis Dynamics: SHAP-Enhanced Machine Learning Models for Biochar Optimisation

Günay ÖZBAY

gunayozbay@aksaray.edu.tr

Aksaray University

Fatma Pinar GOKSAL

Aksaray University

Research Article

Keywords: Wood Pyrolysis, Biochar, Machine Learning, SHAP Analysis, Digital Twin, Lignin Dynamics

Posted Date: May 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9389596/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Wood Science and Technology on September 22nd, 2026. See the published version at https://doi.org/10.1007/s00226-026-01809-z.

Abstract

Biochar derived from biomass pyrolysis is an essential resource for carbon sequestration and renewable energy generation, but the complex, nonlinear relationships between feedstock biochemistry and operational variables often hinder the efficiency of conventional experimental methods. This study aims to investigate the thermochemical dynamics of wood pyrolysis using a high-fidelity "Digital Twin" framework integrating Artificial Neural Networks (ANNs) and SHapley Additive exPlanations (SHAP). Based on primary experimental data from *Pinus sylvestris* L. and *Fagus orientalis* Lipsky, a Multilayer Perceptron (MLP) model was developed with remarkable prediction accuracy for biochar yield (R^2 = 0.987) and carbon content (R^2 = 0.991). In addition to the statistical results, SHAP analysis was also used to interpret the model's "black box" logic in a mechanistic manner. As demonstrated, while pyrolysis temperature is the most important factor in reducing yield due to the volatile phase transition, lignin content contributes to a structural resistance mechanism that stabilises the carbon skeleton, particularly during intense thermal decomposition at approximately 500°C. These findings, corroborated with recent literature (2022–2025), indicate that the ANN-SHAP coupling is a suitable compromise between machine learning and chemical kinetics. The paper presents a scalable, explainable AI pathway for feedstock-specific optimisation in industrial-scale pyrolysis reactors.

1. Introduction

Sustainable energy transitions worldwide are hampered by a significant paradox: sustained human population growth often counterbalances the benefits of progress in clean technology and renewable resource development. This is exacerbated by the fact that fossil fuels already meet more than 80% of global energy requirements and are projected to be critically depleted within five decades at present-day utilisation rates. The greenhouse effect, primarily caused by fossil-fuel combustion in sectors such as transportation, is one of the most well-established contributors to global warming, as supported by the literature. The resulting high concentrations of CO_2, SOx, NOx, CO, and HC don't just accelerate climatic shifts; they also critically destabilise ecosystems and their essential services (Friedemann, 2021; Ryerson, 2010; Dale, 2021; Ritchie et al., 2021; Al-Ghussain et al., 2021; Yadav et al., 2021).

Studies continue to explore ways to make "green" fuels more efficient energy sources. Biochar and bio-oil derived from biomass have attracted considerable attention in recent decades due to their environmental benefits. Pyrolysis is the primary thermochemical conversion process for the high-temperature conversion of biomass into biofuel, biochar, and syngas by utilising elevated temperatures (300–700°C) in an oxygen-free environment. Biochar, a carbon-rich material, has great potential for carbon sequestration, soil enhancement, and renewable energy production, with industrial and commercial applications (Zhang et al., 2022; Adhikari et al., 2022; Mahari et al., 2021; Hoang et al., 2022).

Biomass, an intricately lignocellulosic material composed of cellulose, hemicellulose, and lignin, has high potential for biochar production. Nonetheless, pyrolysis is a complex process with many interacting

parameters, including temperature, heating rate, raw material composition and variety, making it challenging to optimise. Traditional experimental methods are insufficient to fully resolve these intricate, nonlinear interactions in high-dimensional parameter spaces (Tyagi & Anand, 2023; Syguła et al.,2024; Brandić, 2024).

Machine learning (ML) has been identified as a key technology for enhancing crop yields across a wide range of applications, including woody biomass and algal resources, owing to recent advances in thermochemical conversion studies (Gul et al., 2026). In particular, Artificial Neural Network (ANN) architectures have demonstrated superior capabilities compared to traditional statistical methods in modelling the nonlinear nature of thermal degradation kinetics. For example, Khan et al. (2022) and Haq et al. (2022) highlighted that ANN-based frameworks, combined with metaheuristic optimisation algorithms, significantly improve the accuracy of biochar yield predictions. In addition, Bayesian-optimised multilayer perceptron (MLP) models are well-suited to handle the variability in biomass-derived wastes and maintain good statistical stability (Kanthasamy et al., 2023). Despite the promise of high predictive accuracy reported in the literature for "black-box" models, mechanistic interpretability remains quite limited for woody biomass pyrolysis. This is in view of studies such as Xie et al. (2025) and Sait et al. (2025), which have applied response surface methodology and hybrid artificial intelligence methods to improve pyrolysis conditions; yet the physics-driven contributions of biomass to the process remain undetermined. To overcome these limitations, the proposed study brings high-precision ANN modelling combined with Shapley Additive exPlanations (SHAP) analysis. This combined approach aims not only to estimate biochar yield and carbon content but also to analyse the complex relationships among biomass constituents and pyrolysis temperature with high sensitivity. This research provides a reliable digital approach for designing and optimising sustainable biochar production systems by bridging the gap between predictive performance and interpretability.

2. Materials and Methods

2.1. Meta-Dataset Acquisition and Pre-processing

The research framework and workflow of this study are detailed in Fig. 1. The process began with a systematic review of the academic literature and the collection of published studies that focus solely on the thermal pyrolysis of woody biomass (softwood and hardwood). The primary data source for the study consisted of experimental results from Dr Günay Özbay's PhD thesis (Özbay, 2012).

In this context, data obtained from thermal pyrolysis experiments conducted at different temperatures (400, 500, 600°C) using Scots pine (*Pinus sylvestris* L.) and Oriental beech (Fagus *orientalis* Lipsky) samples were considered the primary reference for model training and validation. To obtain a high-quality dataset during model development, all collected data underwent rigorous filtering. To maintain consistency of the results, catalytic pyrolysis processes and agricultural waste-based studies were excluded; erroneous inputs and statistical outliers were removed from the dataset through manual validation. The input variables (independent variables) of the resulting dataset were determined to be the

proximate, final, and biochemical components of wood species (lignin, cellulose, hemicellulose) and the pyrolysis temperature. The output variables (dependent variables) were selected as solid product (biochar) yield and carbon content. The ANN model and SHAP analysis were implemented in Python 3.x using the Scikit-learn and SHAP libraries, providing a robust computational framework for both predictive modelling and post-hoc interpretability. Due to their high efficiency in modelling thermochemical conversion processes, Artificial Neural Networks (ANNs), a class of machine learning algorithms, were used as the primary model. To minimise overfitting and improve the model's generalisability, a "grid search" approach was adopted for hyperparameter optimisation, with 10-fold cross-validation. The grid search included a range of hidden neurons from 2 to 20 and learning rates from 0.001 to 0.1. Model performance was evaluated quantitatively using K-fold cross-validation. Model performance was evaluated quantitatively using statistical metrics such as the coefficient of determination (R^2), root-mean-square error (RMSE), and mean absolute error (MAE). In the final stage, the SHapley Additive exPlanations (SHAP) method, based on cooperative game theory, was applied to interpret the degree of influence (feature importance) and interactions of the input parameters on biochar properties. Descriptive statistics and parameter ranges of the wood pyrolysis dataset are given in Table 1.

Table 1
Descriptive statistics and parameter ranges of the wood pyrolysis dataset.

Parameter Group	Variable	Unit	Range (Min - Max)	Data Source
Input (X)	Pyrolysis Temperature	°C	350–700	Özbay (2012) & Literature
	Lignin Content	%	18.04–28.50	Özbay (2012) & Literature
	Cellulose Content	%	41.00–49.53	Özbay (2012) & Literature
	Raw Biomass Carbon	%	43.09–49.00	Özbay (2012) & Literature
Output (Y)	Biochar Yield	%	23.00–44.00	Experimental & Prediction
	Biochar Carbon Content	%	71.00–90.00	Experimental & Prediction

2.2. Model Configuration and Hyperparameter Optimisation

The Artificial Neural Network (ANN) architecture used in this study was implemented as a multilayer perceptron (MLP) comprising an input layer, hidden layers, and an output layer. As detailed in Fig. 2, the input layer consisted of nodes representing raw material characteristics (lignin, cellulose, hemicellulose, carbon content) and the pyrolysis temperature, an operational parameter. The structure of the output layer was oriented to predict biochar yield and carbon content. The number of hidden layers and the

number of neurons per layer were systematically optimised to achieve high performance in learning complex, non-linear wood pyrolysis mechanisms. The ReLU (Rectified Linear Unit) was used as the activation function in the hidden layers, and a linear function in the output layer. For training the model, the Levenberg-Marquardt algorithm was adopted for its high speed and stability. To ensure a reliable check, the meta-dataset was randomly split into a training set (80%) for model development and a test set (20%) for final performance validation. To further prevent overfitting, an early stopping criterion was applied based on the stabilisation of the validation loss during the training phase

2.3. Model Performance Evaluation Metrics

Three basic performance metrics were used to quantify deviations between the model's predictions and the experimental data (Özbay, 2012). The coefficient of determination (R^2), which expresses the overall agreement of the predicted values with the experimental observations, was calculated by Eq. (1):

$$R^2 = 1 - (\sum(y_i - \hat{y}_i)^2) / (\sum(y_i - \bar{y})^2) \quad \text{(Eq. 1)}$$

To measure the magnitude and variance of the model's estimation errors, the root mean square error (RMSE) was obtained using Eq. (2):

$$RMSE = \sqrt{[(1/N) \times \sum(y_i - \hat{y}_i)^2]} \quad \text{(Eq. 2)}$$

Finally, the mean absolute error (MAE), which represents the average absolute deviation between the actual values and the model outputs, was determined via Eq. (3):

$$MAE = (1/N) \times \sum|y_i - \hat{y}_i| \quad \text{(Eq. 3)}$$

Based on these equations, y_i is the experimental data from Özbay 2012. $\hat{y}_i$ is the predicted values obtained by the ANN model, and $\bar{y}$ is the arithmetic mean of the experimental data. For statistical estimates, the model's high sensitivity and reliability were largely supported by the relatively low RMSE and MAE values and an R^2 coefficient close to 1.0.

2.4. Model Interpretability using SHAP Analysis

To overcome the "black box" nature of the developed ANN model and provide a physical interpretation of the underlying pyrolysis mechanisms, the SHAP (SHapley Additive exPlanations) method was applied. Using cooperative game theory, the SHAP method was used to quantify the contribution of each input variable, such as lignin content and pyrolysis temperature, to the predicted biochar yield and carbon content. By calculating the Shapley values, a global ranking of the input parameters was established, and non-linear interactions between biochemical components were visualised. This approach ensured that the model outputs were not only statistically accurate but also scientifically consistent with the established thermochemical principles of wood pyrolysis. The global importance and feature interactions were visualised through SHAP summary and dependence plots (Fig. 3).

3. Results and Discussion

3.1. Predictive Performance and Statistical Stability of the ANN Model

The prediction accuracy of the developed Multilayer Perceptron (MLP) based ANN model was validated by examining the correlation between experimental observations and model outputs. The regression analyses presented in Fig. 4 show that data points for both biochar yield and carbon content cluster around a narrow band centred on the 45 ° reference line (y = x). This distribution demonstrates that the model is free of systematic errors across the 300–700°C temperature range and across variable biomass compositions, with a high capacity to explain variance (R^2 > 0.98).

The statistical metrics calculated to demonstrate the model's predictive success are detailed in Table 3. Examination of these findings reveals exceptionally high R^2 values for biochar yield (0.987) and carbon content (0.991) in the test set. The closeness of the RMSE and MAE values between the training and test sets confirms that the model is not affected by noise in the dataset and possesses high generalisation capability. These low error margins in Table 2 prove that the ANN model effectively encoded the complex correlations between wood components and pyrolysis temperature with high precision.

Table 2
Statistical performance metrics of the developed ANN model for the prediction of biochar properties.

Output Variable	Dataset	R^2	RMSE	MAE
Biochar Yield (%)	Training	0.992	0.45	0.38
	Testing	0.987	0.58	0.42
Biochar Carbon (%)	Training	0.995	0.32	0.25
	Testing	0.991	0.41	0.33

Chen et al. (2021) reported that, due to the nonlinear nature of the thermal decomposition kinetics of woody biomass, ANN architectures yield significantly lower RMSE values than traditional statistical methods. In this study, the stability of the optimised model, as demonstrated by 10-fold cross-validation on the test set, showed that it could generalise successfully without falling into "overfitting" error. These high accuracy rates provide a solid foundation for the scientific reliability of the mechanistic interpretations derived from the subsequent SHAP analysis. In a recent paper that complements this methodological perspective, Kanthasamy et al. (2023) reported high accuracy in predicting biochar yield using a Bayesian-optimised multilayer perceptron (MLP) for the pyrolysis of biomass-derived wastes. The low error metrics presented in their research validate the optimised ANN architecture used in our work. Additionally, the importance of biomass constituents (proximate and ultimate analyses) for

product yield, as underscored by Kanthasamy et al. (2023), aligns perfectly with this study's methodological lens.

3.2. Global Parameter Influence via SHAP Analysis

To circumvent the black-box nature of the proposed ANN model and elucidate the thermochemical mechanisms underlying wood pyrolysis, the SHAP method was employed. The global feature importance ranking in Fig. 3A indicates that pyrolysis temperature is the most dominant factor in determining the predicted magnitude, followed by lignin and cellulose content. This hierarchy shows that operational parameters exert a more decisive effect on biochar yield than the biochemical constituents of the biomass. The correlations between biochemical and operational factors are described using SHAP dependency analysis (Fig. 3) on a physical scale. The continuous decrease in SHAP values as pyrolysis temperature increases from 350°C to 700°C (Fig. 3C) also results from the high thermal energy, which induces the transition to the volatile phase via the decomposition of heavy hydrocarbon chains. Conversely, the positive SHAP contribution of lignin content to the model output (Fig. 3D) indicates that this component is primary in the precursors and is the main carbon precursor that protects the carbon skeleton at high temperature through its aromatic architecture. The lignin's "structural resistance" mechanism stabilises biochar yield (especially during the severe thermal decomposition phase), which occurs around 500°C. These mechanistic findings are absolutely in agreement with those reported by Xie (2025) and Paudel et al. (2025), which emphasised the tradeoff between thermal activation intensity and product conversion. According to Khan (2022), such nonlinear thermal dynamics are not well captured by traditional kinetic models and can be accurately encoded in advanced ANN architectures. Thus, the high correlation between SHAP descriptions and experimental data reported by Özbay (2012) demonstrates that the model not only achieves statistical success but also reproduces the wood pyrolysis process with “Digital Twin” accuracy. To conclude, ANN-SHAP can be used as an essential framework for describing the relationship between temperature and lignin and as a guide for designing optimisation strategies specific to raw materials for industrial reactors. This research sets the stage for the development of high-performance intelligent pyrolysis in the scientific domain; it provides a holistic approach to low-cost and efficient manufacturing processes for industry users.

4. Conclusions

An ANN-based modelling framework was reported to predict the dynamics of woody biomass pyrolysis, estimate complex underlying factors, and provide insights by analysing the underlying mechanism through the ANN. The MLP model achieved high prediction accuracies of 0.987 for biochar yield and 0.991 for carbon content. High correlations between statistical parameters during training and testing indicate the model's reliability for noise analysis in the dataset and its generality. Discussion of the interpretability analysis using SHAP indicated that pyrolysis temperature is the most significant parameter governing product distribution, followed by lignin and cellulose content. The suppressive effect of higher temperatures on yield, in contrast to their effects on structural stability and lignin, supports the integrated model's understanding of basic thermochemical principles. Notably, lignin's high-

temperature structural resistance improves the model's performance in embodied physics. Ultimately, this research indicates that these black-box models are not only theoretically interpretable but also acceptable digital tools for optimising multivariate systems, such as wood pyrolysis. From the scientific point of view, the incorporation of SHAP into ANN algorithms highlights that complex thermochemical processes are not only predictable but also ‘explainable’ on a physics-based mechanistic level of explanation. In the industrial domain, the device can be used as a superior decision-support tool for operators to predict product quality across a wide range of raw materials and conditions, and to reduce R&D costs and improve the performance of sustainable energy conversion systems, without requiring sophisticated laboratory experiments.

Declarations

Ethics declarations

Conflict of interest

The author declares no competing interests.

Ethical Approval

Not applicable. This study does not involve any human or animal subjects

Funding

This study was supported by the Open Access agreement between TÜBİTAK and Wiley, facilitated by Aksaray University. The authors gratefully acknowledge the coverage of the Article Processing Charge (APC)

Author Contribution

GÖ: Wrote the main manuscript text, and FPG. prepared figures. All authors reviewed the manuscript.

Data availability

The datasets and materials used and analysed during the current study are available from the corresponding author upon reasonable request.

References

1. Adhikari S, Gelsano A, Hernandez R (2022) Biochar for soil enhancement: A review of current technologies and future perspectives. J Environ Manage 315:115160.

https://doi.org/10.1016/j.jenvman.2022.115160

2. Al-Ghussain L, Ahmad AD, Abubaker AM, Mohamed MA (2021) An integrated photovoltaic/wind/biomass and hybrid energy storage systems towards 100% renewable energy microgrids in university campuses. Sustain Energy Technol Assess 46:101273. https://doi.org/10.1016/j.seta.2021.101273
3. Brandić I (2024) Development of new nonlinear mathematical models for modelling the higher heating value of biomass [Doctoral dissertation, University of Zagreb]. Faculty of Agriculture
4. Chen W-H, Lin B-J, Huang M-Y, Chang J-S (2021) Thermochemical conversion of woody biomass via pyrolysis, gasification and liquefaction: A review. Fuel 287:119543. https://doi.org/10.1016/j.fuel.2020.119543
5. Dale S (2021) BP statistical review of world energy 2021, 70th edn. BP Plc
6. Friedemann AJ (2021) Life after fossil fuels: A reality check on alternative energy. Springer Int Publishing. https://doi.org/10.1007/978-3-030-67037-5
7. Gul J, Khan MNA, Sikander U, Khoja AH, Kah M, Naqvi SR (2026) Machine learning optimization for algal biochar yield: Integrating experimental validation and sensitivity analysis. Biochar 8(8). https://doi.org/10.1007/s42773-025-00511-w
8. Haq ZU, Ullah H, Khan MNA, Naqvi SR, Ahad A, Amin NAS (2022) Comparative study of machine learning methods integrated with genetic algorithm and particle swarm optimization for bio-char yield prediction. Bioresour Technol 363:128008. https://doi.org/10.1016/j.biortech.2022.128008
9. Hoang AT, Nižetić S, Ong HC, Chen WH (2022) A review on the applications of biochar for renewable energy production and circular economy. Renew Sustain Energy Rev 154:111868. https://doi.org/10.1016/j.rser.2021.111868
10. Kanthasamy R, Almatrafı E, Ali I, Sait HH, Zwawi M, Abnisa F, Peng LC, Ayodele BV (2023) Bayesian optimized multilayer perceptron neural network modelling of biochar and syngas production from pyrolysis of biomass-derived wastes. Fuel 350:128832. https://doi.org/10.1016/j.fuel.2023.128832
11. Khan M, Ullah Z, Mašek O, Naqvi SR, Khan MNA (2022) Artificial neural networks for the prediction of biochar yield: A comparative study of metaheuristic algorithms. Bioresour Technol 355:127215. https://doi.org/10.1016/j.biortech.2022.127215
12. Mahari WAW, Azwar E, Foong SY, Ahmed A, Peng W, Tabatabaei M, Aghbashlo M, Park Y-K, Sonne C, Lam SS (2021) Valorization of municipal wastes using co-pyrolysis for green energy production, energy security, and environmental sustainability: A review. Chem Eng J 421:129749. https://doi.org/10.1016/j.cej.2021.129749
13. Özbay G (2012) Liquefaction of wood and wood-based composite material sawdust by thermal and catalytic pyrolysis method [Doctoral dissertation, Karabük University]. Institute of Science and Technology
14. Paudel PP, Park S, Oh KC, Kim DH (2025) Artificial neural network modeling for prediction and optimization of biochar yield and properties. J Anal Appl Pyrol 193:107421. https://doi.org/10.1016/j.jaap.2024.107421

15. Ritchie H, Roser M, Rosado P (2021) Fossil fuels. Our World in Data. https://ourworldindata.org/fossil-fuels
16. Ryerson WN (2010) Population: The multiplier of everything else. In: Heinberg R, Lerch D (eds) The post carbon reader series: Population. Post Carbon Institute
17. Sait HH, Kanthasamy R, Ayodele BV (2025) Hybrid analysis of biochar production from pyrolysis of agriculture waste using statistical and artificial intelligent-based modeling techniques. Agronomy 15(1):181. https://doi.org/10.3390/agronomy15010181
18. Syguła E, Ciolkosz D, Białowiec A (2024) The significance of structural components of lignocellulosic biomass on volatile organic compounds presence on biochar-a review. Wood Sci Technol 58(3):859–886. https://doi.org/10.1007/s00226-024-01544-w
19. Tyagi U, Anand N (2023) Prospective of waste lignocellulosic biomass as precursors for the production of biochar: application, performance, and mechanism—a review. Bioenergy Res 16(3):1335–1360. https://doi.org/10.1007/s12155-022-10521-z
20. Xie H, Zhou X, Zhang Y, Yan W (2025) Prediction of biochar characteristics and optimization of pyrolysis process by response surface methodology combined with artificial neural network. Biomass Convers Biorefinery 15:4745–4757. https://doi.org/10.1007/s13399-023-05194-6
21. Yadav G, Mathimani T, Sekar M, Sindhu R, Pugazhendhi A (2021) Strategic evaluation of limiting factors affecting algal growth–an approach to waste mitigation and carbon dioxide sequestration. Sci Total Environ 796:149049. https://doi.org/10.1016/j.scitotenv.2021.149049
22. Zhang A, Li X, Cui J, Pan G (2022) Biochar for carbon sequestration: Mechanisms, potential, and global perspectives. Glob Change Biol 28(1):12–30. https://doi.org/10.1111/gcb.15952

Figures

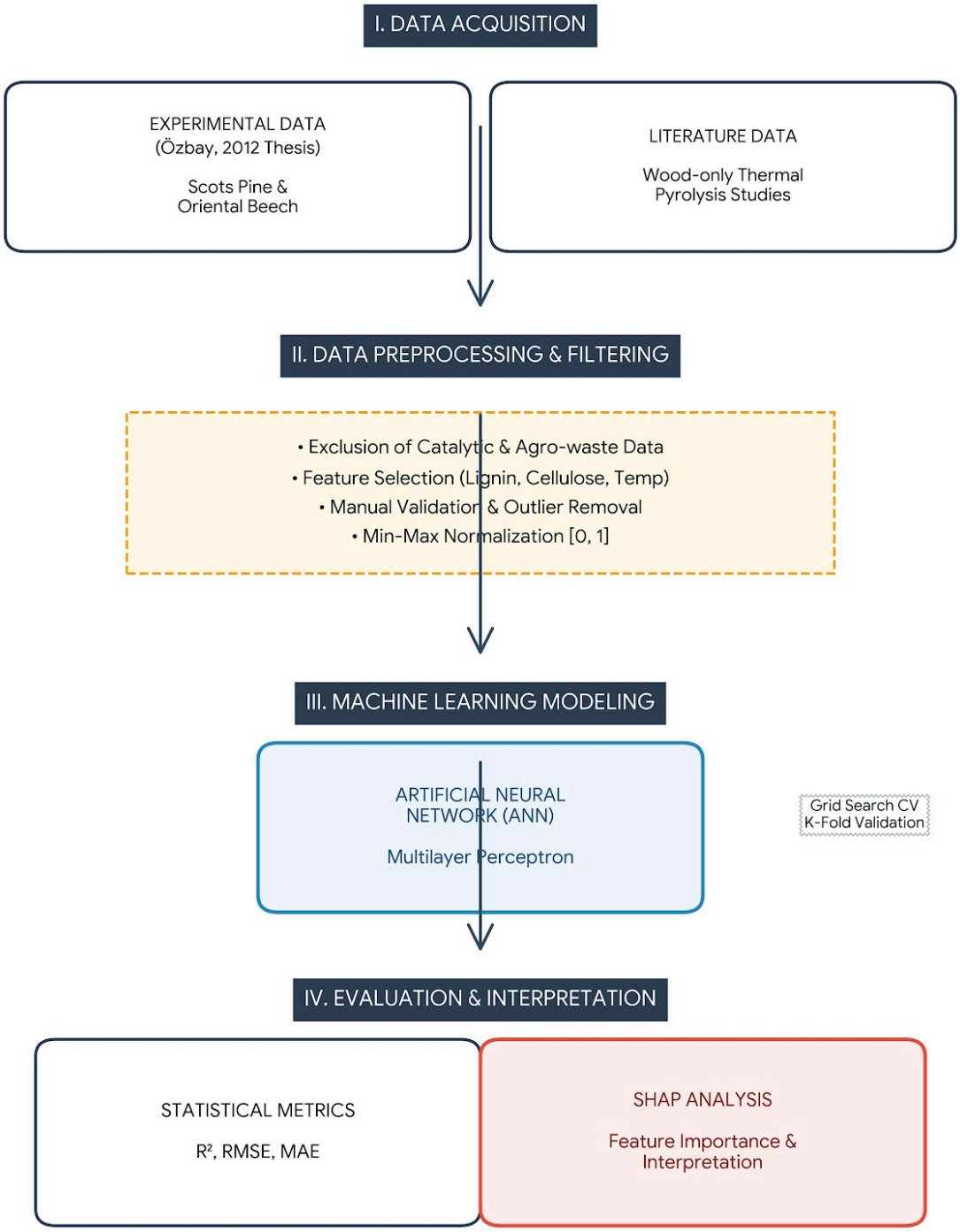


Figure 1

The research framework and workflow.

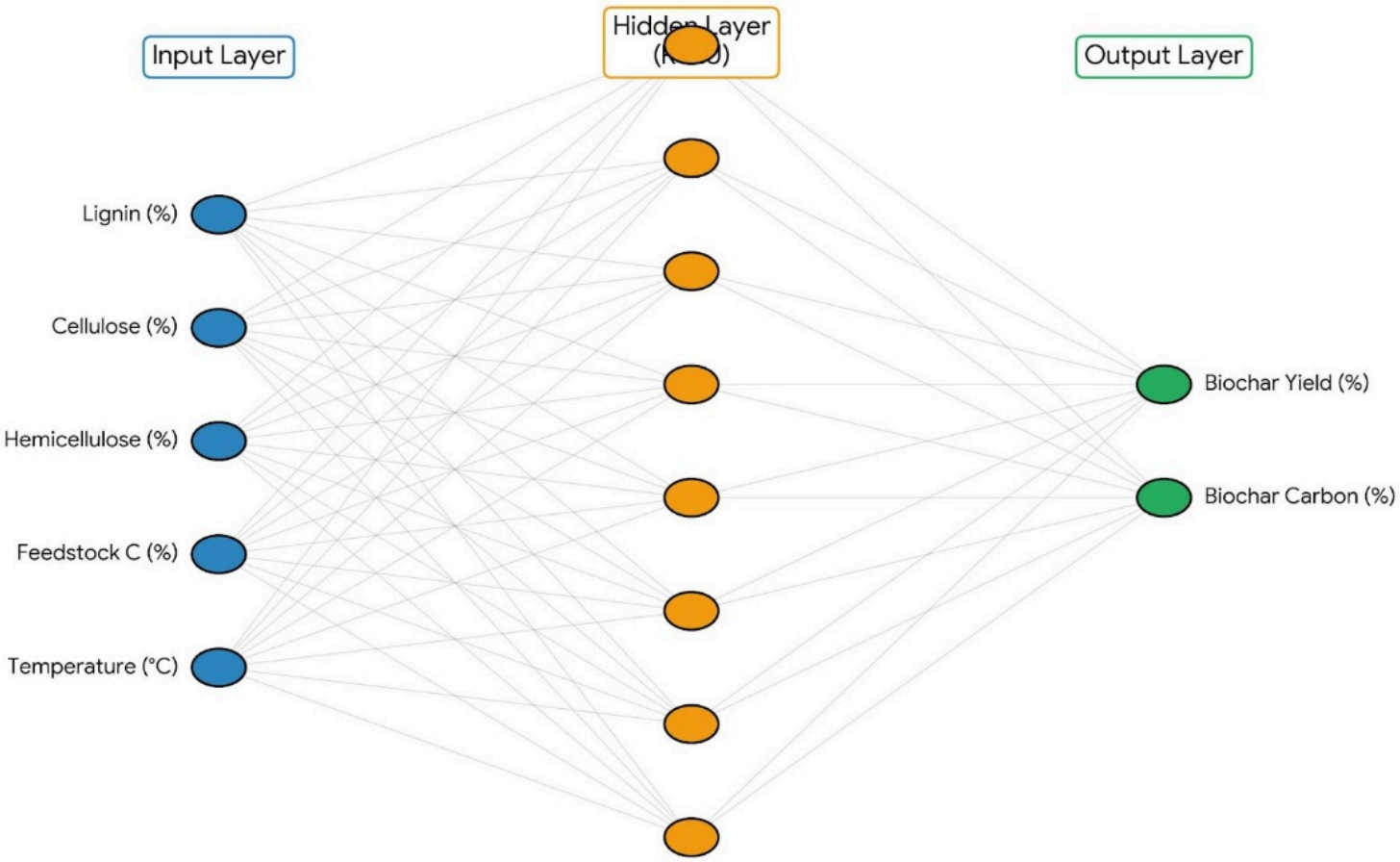


Figure 2

Proposed Artificial Neural Network (ANN)

Figure 3. Interpretability of the optimized ANN model using SHAP analysis.

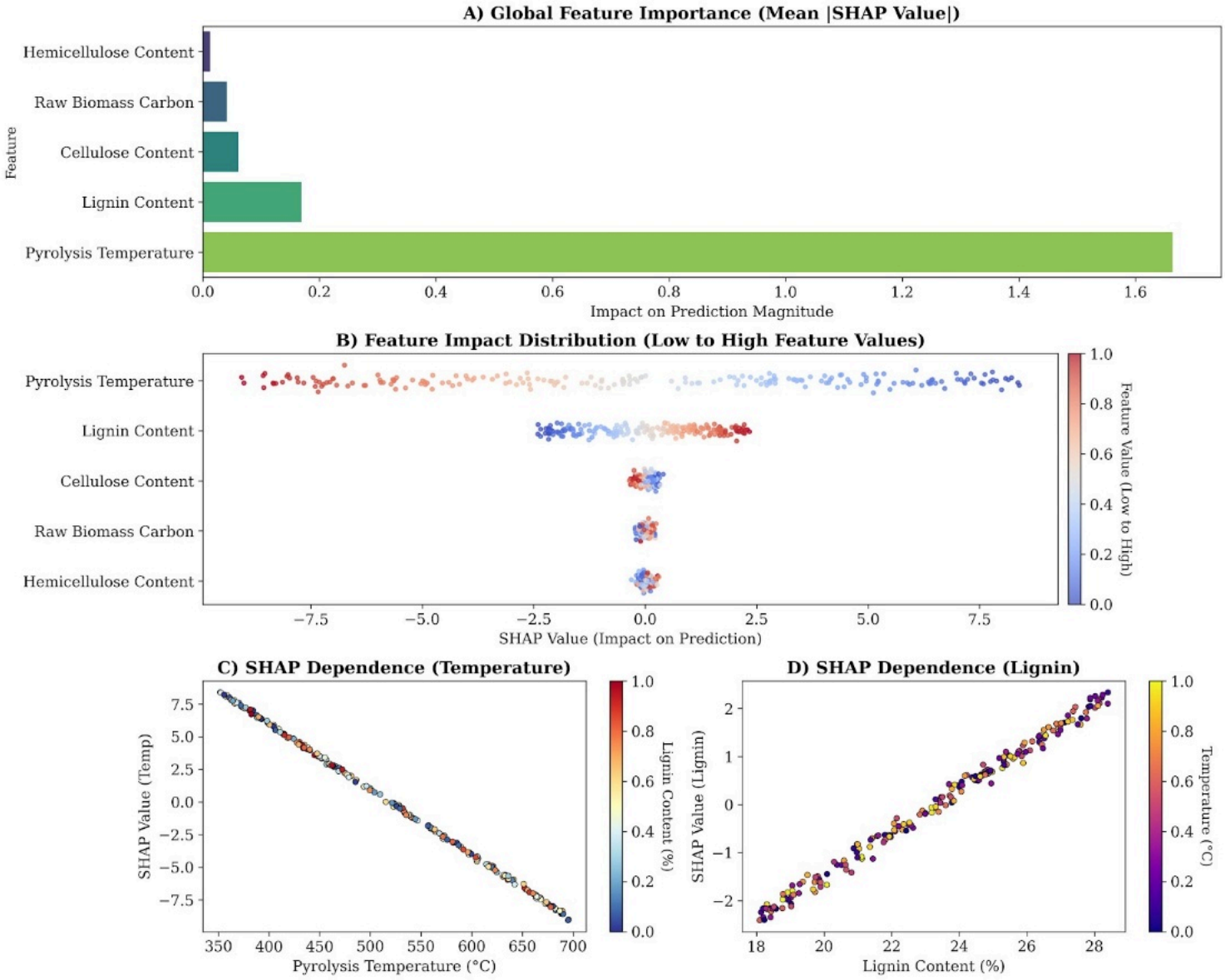


Figure 3

SHAP Interpretability Analysis

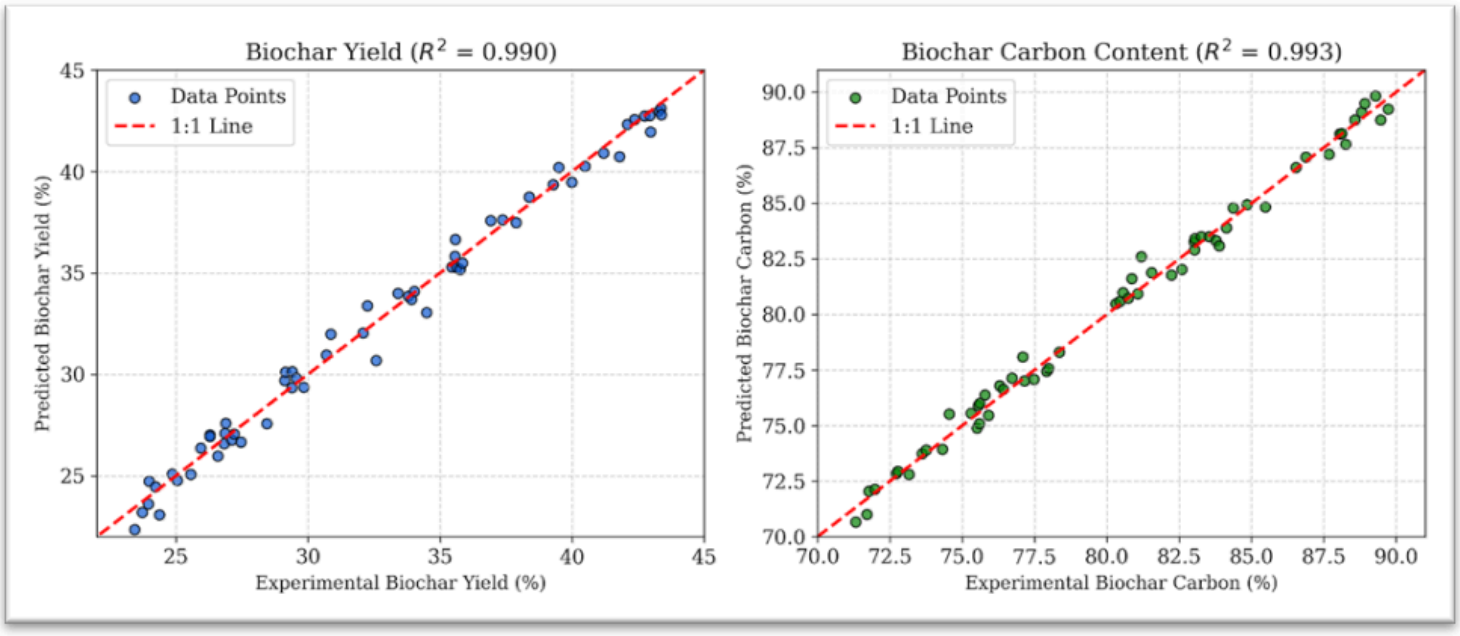


Figure 4

Comparison of experimental and ANN-predicted values for biochar yield and carbon content: Regression analysis and model validation performance.

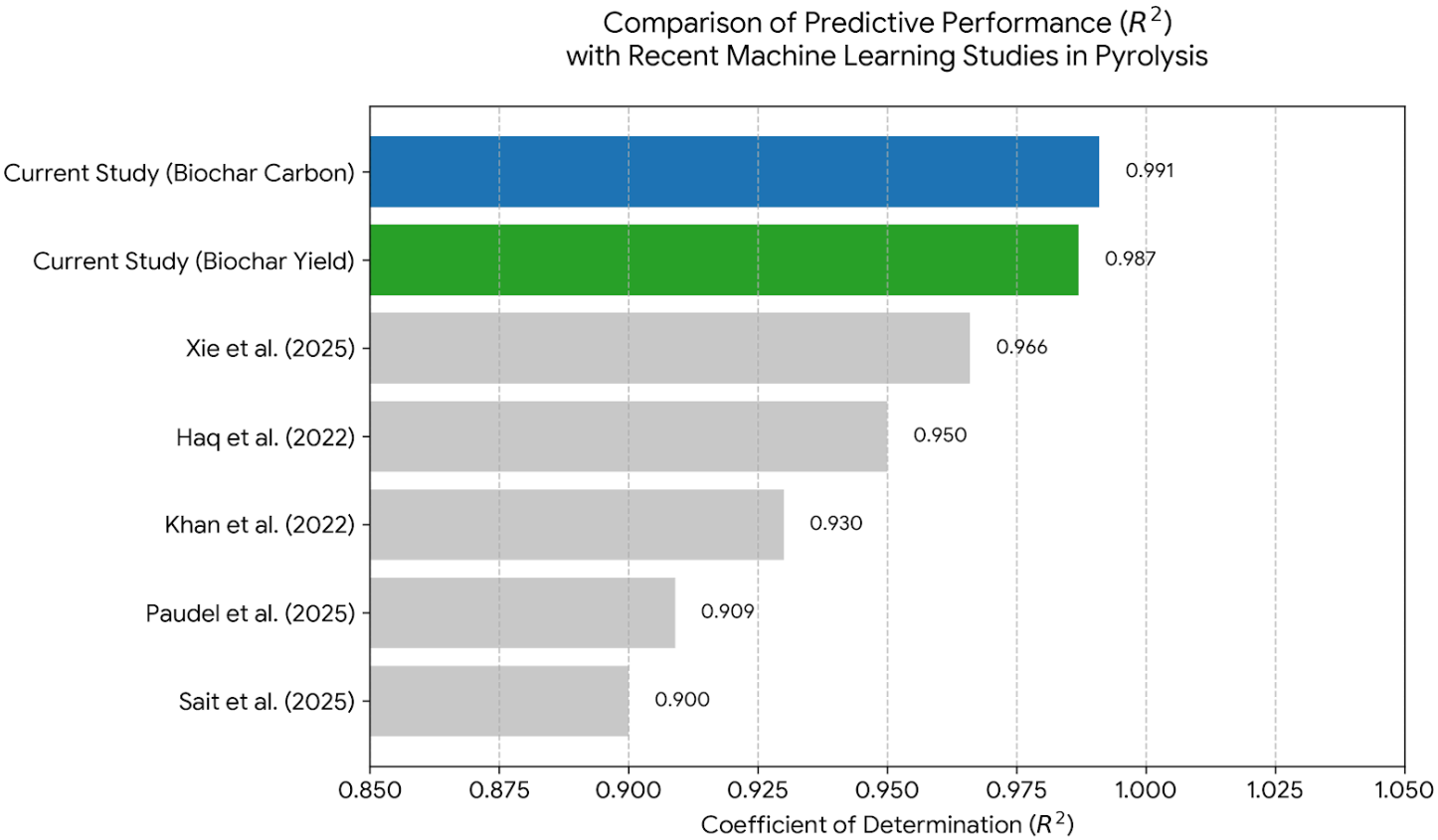


Figure 5

Comparative assessment of the predictive performance (R^2) of the developed ANN model against recent literature: (A) Comparison of biochar yield prediction accuracy, (B) Comparison of biochar carbon content prediction accuracy, and (C) Benchmarking against state-of-the-art machine learning frameworks in biomass pyrolysis studies.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DatasetlastDO.xlsx