CHEMICAL ENGINEERING TRANSACTIONS
VOL. 119, 2025
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: David Bogle, Flavio Manenti, Piero Salatino

ISBN 979-12-81206-20-5; **ISSN** 2283-9216
DOI: 10.3303/CET25119084

Predicting Bio-oil Yield from Biomass Pyrolysis using Machine Learning-Based Tools

Antonio Coppola[a*], Antonio Elia Pascarella[b], Stefano Marrone[b], Roberto Chirone[c], Carlo Sansone[b], Piero Salatino[c]

a STEMS, Consiglio Nazionale delle Ricerche, 80125 Napoli, Italy
b DIETI, Università degli Studi di Napoli Federico II, 80125 Napoli, Italy
c DICMaPI, Università degli Studi di Napoli Federico II, 80125 Napoli, Italy
antonio.coppola@stems.cnr.it

Biomass pyrolysis involves the thermal decomposition of biomass constituents to yield valuable compounds to be exploited as biofuels and/or platform chemicals. Modelling biomass pyrolysis is challenging, but modern AI methods may give a valuable contribution to prediction of process yields, provided high-quality datasets are available. The published literature on the subject mostly refers to limited datasets, usually only a few hundred records, which are inadequate for robust AI applications.
This work presents a dataset of about 500 observations with no missing values, compiled from published data on bio-oil/bio-liquid production via fixed-bed pyrolysis of different biomass. The dataset includes physicochemical properties of the biomass, key pyrolysis operating conditions, and bio-liquid yield. Each observation was carefully standardized to resolve inconsistencies in the terminology and/or lack of standardization. Best results, obtained from XGBoost, showed a MAE of 2.0 and an R^2 of 0.8. Critical analysis of results demonstrates that AI applied to biomass pyrolysis data displays very good predictive ability. However, the ability to reproduce known relationships among key variables of the biomass and of the process appears to be more problematic. This is well shown by analysis of PDP plots, where some inconsistencies with known trends emerge when assessing the influence of selected variables on bio-liquid yield. Moreover, pronounced discrepancies with previously published studies by other research groups are observed when analyzing directional trends.

1. Introduction

The energy sector is experiencing the beneficial effects of digitalization, which is bound to play an increaslingly important role in the development of future multi-sector/multi-vector energy systems. According to the IEA, Artificial intelligence (AI) is stemming out as a powerful tool to accelerate the transition toward more interconnected, reliable, and sustainable energy systems.The expanding utilization of renewable energy sources will help address future energy challenges, but large-scale deployment introduces uncertainties, related to lack of predictability and dispatchability, that may affect system stability and robustness. The availability of reliable tools for energy production forecasting may mitigate these uncertainties. Data-driven techniques are particularly effective, as they allow the identification of complex relationships between variables without requiring the development of complex, sometimes unmanageable, mathematical models.
AI is revolutionizing many industries, and Machine Learning (ML) – a branch of AI – is emerging as a promising tool to manage uncertainties in the renewable energy sector. ML utilizes mathematical models and algorithms to extract information, recognize patterns, and predict complex outcomes (World Economic Forum, 2021). These techniques are being applied to various aspects of renewable energy, including design, optimization, management, distribution, and policymaking (Jha et al., 2017). However, their use is still limited to specific fields, and most research focuses on forecasting solar and wind energy production, leaving significant potential for other energy sources (Lai et al., 2020).

Paper Received: 11 June 2025; Revised: 11 July 2025; Accepted: 30 October 2025
Please cite this article as: Coppola A., Pascarella A.E., Marrone S., Chirone R., Sansone C., Salatino P., 2025, Predicting Bio-oil Yield from Biomass Pyrolysis Using Machine Learning-based Tools, Chemical Engineering Transactions, 119, 499-504 DOI:10.3303/CET25119084

Among renewables, bioenergy has steadily kept a central role due to its dispatchability and associated environmental benefits, especially when implemented in waste-to-energy schemes (Wang et al., 2017). Moreover, bioenergy is comparatively less dependent on availability and use of critical raw materials (IEA). Thermochemical conversion provides efficient, cost-effective and fairly feedstock-agnostic transformation paths of biomass into biofuels (Kang et al., 2021). Pyrolysis, conducted in an oxygen-free or oxygen-limited environment, produces bio-oil, biochar, and non-condensable gases. If properly upgraded, bio-oil could serve as an alternative to fossil fuels. Bio-oil properties depend on both the biomass type and pyrolysis conditions. In this study, the potential of Machine Learning to predict the characteristics of final products and to optimize plant operations is investigated.

Artificial neural networks have been applied to simplify complex kinetic models of pyrolysis, significantly reducing computational costs while maintaining prediction accuracy. This facilitates the integration of detailed kinetic models into broader simulations, optimizing the pyrolysis process on an industrial scale. Additionally, ML has been used alongside experimental data to enhance the pyrolysis parameters of sewage sludge and to model other thermochemical processes, such as gasification.

Research on ML applications in pyrolysis mainly focuses on predicting product yields, particularly bio-oil, followed by biochar and, less frequently, biogas. Data are primarily obtained from fixed-bed pyrolyzers, although there is increasing interest in fluidized-bed systems. Most of the studied biomass is lignocellulosic, and the parameters used for predictions include ultimate and proximate analyses, pyrolysis temperature, heating rate, and biomass particle size.

The aim of this work is to apply machine learning techniques to the pyrolysis process, for the evaluation of the bioliquid yield, using a larger dataset than that made by other research groups. Furthermore, in addition to evaluating the prediction performances of different ML techniques, particular attention was paid to any inconsistencies of these techniques between the influence of the different variables on the bioliquid yield and on the chemical-physical mechanisms relevant for pyrolysis.

2. Methods

2.1 Data Collection

Table 1: Mean, Standard Deviation, min and max values and quartiles of the collected data.

	Mean	Minimum value	First quartile	Second quartile (median)	Third quartile	Maximum value	Standard deviation
Ash(wt%)	5.60	0.11	2.33	5.68	7.29	40.08	3.82
FixedCarbon(wt%)	15.86	0.11	11.98	14.88	17.11	78.55	8.04
Volatiles(wt%)	78.54	10.86	75.79	78.85	83.00	95.98	9.36
C(wt%)	49.21	19.49	44.82	48.50	52.70	79.77	6.51
H(wt%)	6.50	2.41	5.89	6.23	6.74	10.59	1.14
O(wt%)	40.78	10.49	34.53	41.63	48.03	54.12	8.42
N(wt%)	2.89	0.17	0.87	1.87	4.40	22.50	2.66
Cellulose(wt%)	34.45	5.75	27.20	32.49	43.00	60.62	11.47
Hemicellulose(wt%)	27.81	3.40	19.40	25.52	36.55	51.34	10.82
Lignin(wt%)	22.80	0.80	15.00	26.11	30.10	50.40	11.18
T(°C)	513.70	300.00	450.00	500.00	550.00	900.00	89.64
HeatingRate(°C/min)	66.60	5.00	7.00	20.00	50.00	800.00	116.10
ParticleSize(mm)	0.80	0.10	0.45	0.64	1.00	10.00	0.58
FlowRate-Nitrogen (ml/min)	117.95	0.00	0.00	100.00	100.00	2000.00	216.35
Bio Liquid yield (wt%)	40.63	11.00	33.40	40.75	47.83	80.70	9.88
O-Biooil (wt%)	25.29	8.50	19.65	25.21	29.35	49.28	8.01
H-biooil (wt%)	8.37	1.85	7.30	8.24	9.03	12.10	1.49
Aqueous phase (wt%)	17.11	3.89	11.60	14.44	23.02	40.01	7.37

The dataset used was composed from the combination of 4 different datasets already available in the literature from different research groups (Marianela Ortiz, 2021; Tang et al., 2020; Ullah et al., 2021; Zhang et al., 2022) plus an additional extension carried out by the authors for a total of more than 1000 observations for tests regarding biomass pyrolysis in fixed bed reactors. 18 different features have been collected for each observation which can be categorized in:

1. biomass properties: such as Proximate analysis (Ash, Volatile matter and Fixed Carbon on dry basis), Ultimate analysis (Carbon, Hydrogen, Nitrogen, and Oxygen by difference), and macro-components (cellulose, hemicellulose, and lignin) all expressed as mass percentages;

2. pyrolysis conditions: temperature (°C), Heating rate (°C/min) and Flow Rate-Nitrogen (ml/min);
3. pyrolysis performance: bio liquid yield, calculated as total condensed product (organic + aqueous phases) and expressed as weight percentage respect to the initial mass of biomass.

The set of variables chosen between proximate, ultimate, macro-components and operating conditions is such that the correlations between the variables are negligible, that is below the threshold of 0.3 in absolute values (Shazeer et al., 2017): Ash, H/Ceff , Cellulose, Lignin, Pyrolysis Temperature, Heating Rate, Particle Size, were selected for this analysis. The bio-liquid yield, calculated as total condensed product (organic + aqueous phases) and expressed as weight percentage respect to the initial mass of biomass, has been considered as main feature for the characterization of pyrolysis in a fixed bed reactor.

2.2 Experimental Setup

The experiments were conducted on a subsample of the original dataset obtained by all the observations without missing values, resulting in 468 observations. Among the regression models used to establish a benchmark on bio-liquid yield are XGBoost (XGB), Multilayer Perceptron (MLP), Support Vector Regressor (SVR), and Mixture of Experts (MoE) and implemented respectively using the libraries xgboost for XGB, scikit-learn for RF, SVR, and MLP and PyTorch for MoE in Python. The experimental setup was designed to evaluate the performance of different machine learning models on the dataset using 10-fold cross-validation (CV). The test set was not utilized during the hyperparameter optimization process in each CV step to avoid data leakage. It was only used once to estimate the model's generalization error. Hyperparameters for each machine learning model were optimized using genetic algorithms to ensure optimal performance; unlike (Ullah et al., 2021), genetic algorithms were used to optimize hyperparameters and not for feature selection. The genetic algorithm was implemented using the "ContinuousGenAlgSolver" class from the general library. The hyperparameters optimized for each model were as follows:

- XGBoost: depth, learning rate, and the number of trees;
- Random forest: depth and number of trees;
- Support vector regressor: epsilon and regularization parameter;
- Multilayer perceptron: number of layers, number of neurons per layer, learning rate, and epochs;
- Mixture of Experts: number of layers, number of neurons per layer, learning rate, epochs and number of experts.

3. Results

Table 2 shows the best performances as mean squared error, root mean squared error, mean absolute error and r-squared respectively of the 5 techniques used; XGB performed better with 17.80±11.54, 4.21±1.21, 2.28±0.58 and 0.80±0.12 for the bio-liquid yield. Figure 1 reports the parity plot of predicted values of the bio-liquid yield versus actual observed values for the XGB regression model. It should be noted that in this context of tabular data with about 500 samples, although not in a proper Big Data context, the most successful methodologies have not been those involving the use of deep models, such as the use of Mixture of Experts as a predictive model, which could require even hundreds of thousands of instances in a dataset to achieve optimum performance. Results indicate that a system based on machine learning can reliably estimate the bio-liquid yields produced by fixed-bed pyrolysis, with fairly good performance metrics even in such a context of a large and heterogeneous dataset, helping to deal with the inherent uncertainties present in the process, due to variable properties of the biomasses and different operating conditions. The PDP plots for the different features that predict the bio-liquid yield are reported in Figure 2. The PDP plot reveals a negative effect on bio-liquid yield with increasing ash content in the biomass.

Table 2. Bio-liquid yield benchmark on the reduced dataset without missing data.

model	mean squared error	root mean squared error	mean absolute error	R^2
MoE	169.93±40.14	13.03±1.46	10.63±1.61	0.00±0.41
XGB	17.80±11.54	4.21±1.21	2.28±0.58	0.80±0.12
MLP	47.49±13.64	6.82±0.97	5.01±0.88	0.48±0.14
SVR	44.33±17.83	6.50±1.42	3.86±0.76	0.52±0.17
RF	23.10±14.09	4.61±1.34	3.07±0.64	0.75±0.15

This trend, also observed by others (Leng et al., 2023; Yang et al., 2022), is consistent with the role of ash as catalyst promoting secondary cracking of pyrolysis vapors reducing liquid yield and quality (Bridgwater, 2012). Conversely, H/C_{eff}, as documented in the pertinent literature (Cheng & Huber, 2012; Degnan Jr, 1986), has a positive effect on the production of bio-liquids, conversely limiting the production of coke. As concerned cellulose

and lignin, these two features have both a negative effect on the production of bio-liquid, with a greater impact of cellulose. The fate of cellulose and lignin during pyrolysis results from the competition between primary decomposition reactions, leading to bio-liquid, and secondary polymerization of vapors, leading to biochar formation and this negative effect in this case could be explained in the light of the prevailingly slow pyrolysis conditions typical of fixed beds, and by the accumulation of biochar (Troiano et al., 2022) with fairly small H/C_{eff} ratio (Li et al., 2023) that may favor the production of further bio-char at the expense of bio-liquid. The effect of the pyrolysis temperature is among the most studied in literature, with observations typically converging toward an indication of 500-550°C as the temperature range to maximize the bio-liquid yield. Like T, also the heating rate is considered one of the key parameters to switch from slow to fast pyrolysis, which involves a greater yield of bio-liquid as its value increases. For both these features the PDP graphs are fully consistent with these observations.

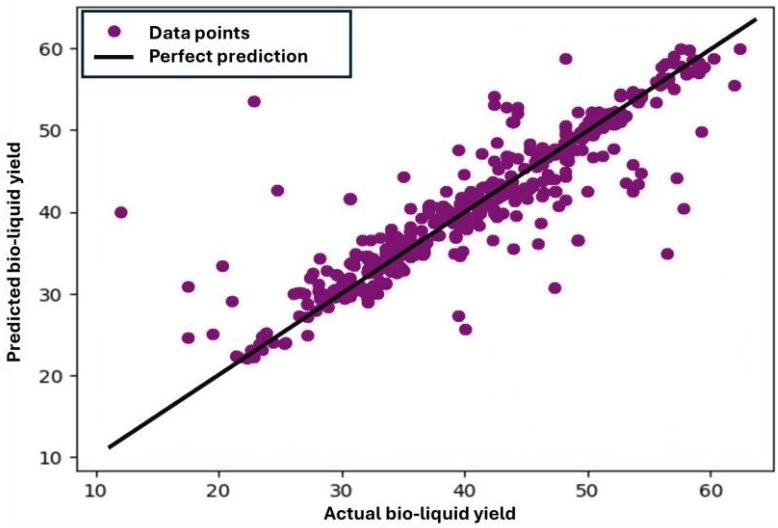

Figure 1: Predicted versus observed bio-liquid yield.

The PDP graph is fully consistent with this observation. This is consistent with indications of the model reported in the PDP graph. On the contrary, the effect of particle size reported by PDP graph discords with respect to the real effect, detecting a positive effect on the bio-liquid yield with the increase of the variable; instead, the increase of the particle size can determine a slow heating rate of the particles with the consequent lowering of the bio-oil yield. This erratic trend is also detected by other research groups (Tang et al., 2020; Ullah et al., 2021; Yang et al., 2022; Zhang et al., 2022).

4. Conclusions

In the present study, five different ML techniques have been used for the forecasting of bio-liquid yield produced by the pyrolysis of residual biomass. The training was conducted on a dataset of about 500 observations without missing data, which is a subset of a more extended database of nearly 1000 observations. Ash content, H/C_{eff}, Cellulose and Lignin contents, Pyrolysis Temperature, Heating Rate, Particle Size, were selected for this analysis and the best regression model was XGBoost, yielding a MAE of 2.28 and an R^2 of 0.80. The analysis of the PDP plots highlighted a partial congruence among the influence of the different variables on the investigated bio-liquid yield and chemical-physical mechanisms relevant to pyrolysis. Also, the comparison of results generated by the various research groups highlighted remarkable discrepancies in directional trends. This criticality reflects a more general caution about using purely data-driven ML for interpretative analysis, and the possibility that ML models lead to predicted trends that are inconsistent with the physical constraints. This finding stimulates further research on the development of physics-informed machine learning tools that embodies physico-chemical relationships and constraints pertinent of the specific domain.

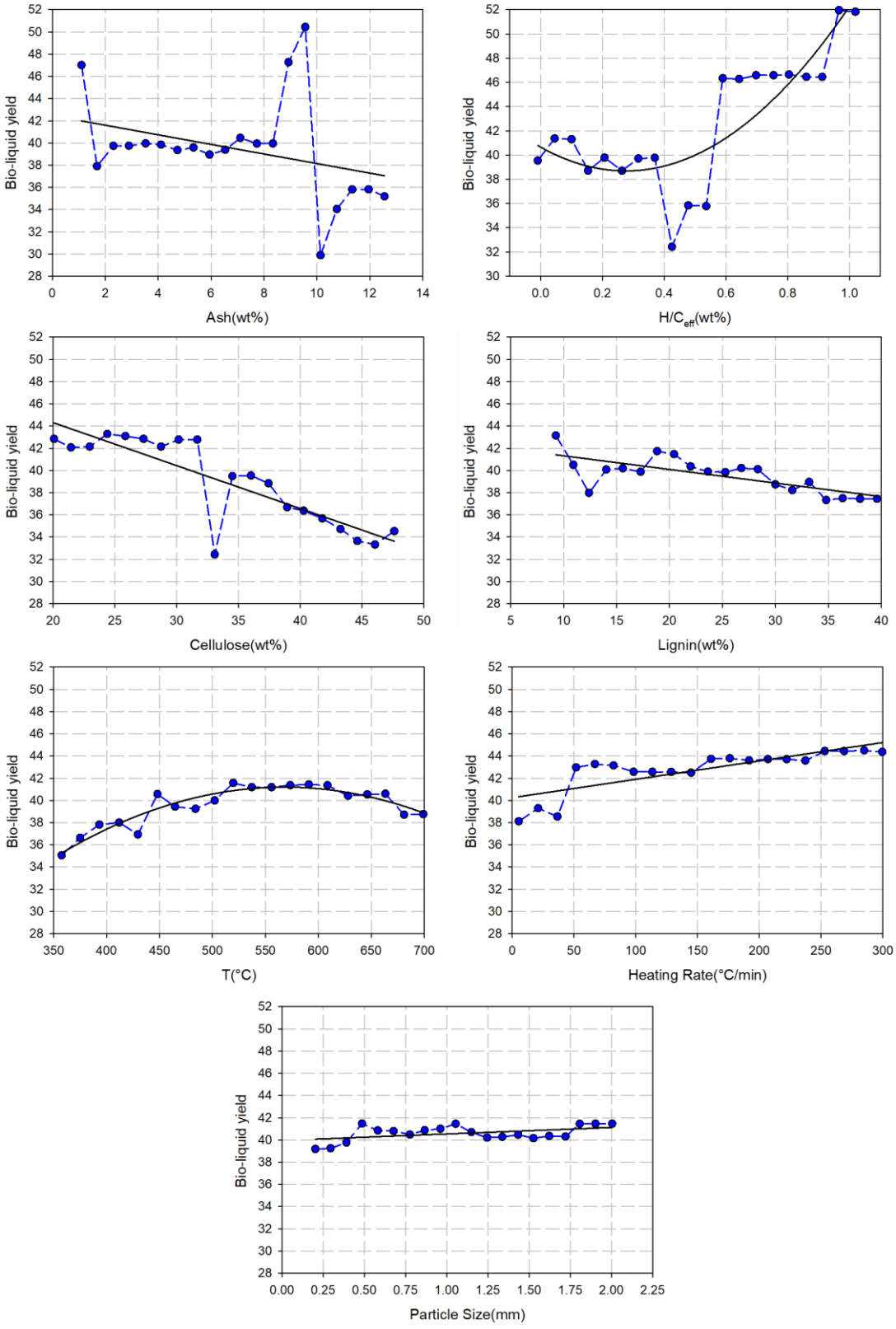

Figure 2. Partial dependency plots (PDP) for bio-liquid yield with respect to biomass properties and operating conditions.

Acknowledgments

Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of 15.03.2022 of Ministero dell'Università e della Ricerca (MUR); funded by the European Union – NextGenerationEU; Project code PE0000021, Concession Decree No. 1561 of 11.10.2022 adopted by Ministero dell'Universita` e della Ricerca (MUR), CUP E63C22002160007 and B53C22004060006, Project title "Network 4 Energy Sustainable Transition – NEST".

References

Bridgwater, A. V. (2012). Review of fast pyrolysis of biomass and product upgrading. Biomass and Bioenergy, 38, 68–94. https://doi.org/10.1016/j.biombioe.2011.01.048

Cheng, Y. T., & Huber, G. W. (2012). Production of targeted aromatics by using Diels-Alder classes of reactions with furans and olefins over ZSM-5. Green Chemistry, 14(11), 3114–3125. https://doi.org/10.1039/c2gc35767d.

Degnan Jr, T. F. (1986). Liquid fuel from carbohydrates. Chemtech, 506–511.

Jha, S. K., Bilalovic, J., Jha, A., Patel, N., & Zhang, H. (2017). Renewable energy: Present research and future scope of Artificial Intelligence. Renewable and Sustainable Energy Reviews, 77(April), 297–317. https://doi.org/10.1016/j.rser.2017.04.018

Kang, K., Klinghoffer, N. B., ElGhamrawy, I., & Berruti, F. (2021). Thermochemical conversion of agroforestry biomass and solid waste using decentralized and mobile systems for renewable energy and products. Renewable and Sustainable Energy Reviews, 149(June), 111372. https://doi.org/10.1016/j.rser.2021.111372

Lai, J. P., Chang, Y. M., Chen, C. H., & Pai, P. F. (2020). A survey of machine learning models in renewable energy predictions. Applied Sciences (Switzerland), 10(17). https://doi.org/10.3390/app10175975

Leng, L., Li, T., Zhan, H., Rizwan, M., Zhang, W., Peng, H., Yang, Z., & Li, H. (2023). Machine learning prediction of the yield and oxygen content of bio-oil via biomass characteristics and pyrolysis conditions. Energy, 278(PB), 127967. https://doi.org/10.1016/j.energy.2023.127967

Li, Y., Gupta, R., Zhang, Q., & You, S. (2023). Review of biochar production via crop residue pyrolysis: Development and perspectives. Bioresource Technology, 369(September 2022), 128423. https://doi.org/10.1016/j.biortech.2022.128423

Marianela Ortiz. (2021). Biomass pyrolysis dataset. Mendeley. https://doi.org/10.17632/BX88YMGBBV.1

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings, 1–19.

Tang, Q., Chen, Y., Yang, H., Liu, M., Xiao, H., Wu, Z., Chen, H., & Naqvi, S. R. (2020). Prediction of Bio-oil Yield and Hydrogen Contents Based on Machine Learning Method: Effect of Biomass Compositions and Pyrolysis Conditions. Energy and Fuels, 34(9), 11050–11060. https://doi.org/10.1021/acs.energyfuels.0c01893

Troiano, M., Ianzito, V., Solimene, R., Ganda, E. T., & Salatino, P. (2022). Fluidized Bed Pyrolysis of Biomass: A Model-Based Assessment of the Relevance of Heterogeneous Secondary Reactions and Char Loading. Energy and Fuels, 36, 9660−9671. https://doi.org/10.1021/acs.energyfuels.2c01483

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

Wang, S., Dai, G., Yang, H., & Luo, Z. (2017). Lignocellulosic biomass pyrolysis mechanism: A state-of-the-art review. Progress in Energy and Combustion Science, 62, 33–86. https://doi.org/10.1016/j.pecs.2017.05.004

World Economic Forum. (2021). Harnessing Artificial Intelligence to Accelerate the Energy Transition. White Paper, September, 25.

Yang, K., Wu, K., & Zhang, H. (2022). Machine learning prediction of the yield and oxygen content of bio-oil via biomass characteristics and pyrolysis conditions. Energy, 254, 124320. https://doi.org/10.1016/j.energy.2022.124320

Zhang, T., Cao, D., Feng, X., Zhu, J., Lu, X., Mu, L., & Qian, H. (2022). Machine learning prediction of bio-oil characteristics quantitatively relating to biomass compositions and pyrolysis conditions. Fuel, 312(September 2021), 122812. https://doi.org/10.1016/j.fuel.2021.122812