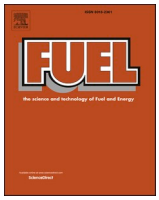
Fuel 394 (2025) 135000

Contents lists available at ScienceDirect

Fuel

journal homepage: www.elsevier.com/locate/fuel



Full Length Article

Critical assessment of machine learning prediction of biomass pyrolysis

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

[a] Department of Electrical Engineering and Information Technology (DIETI), University of Naples Federico II, Via Claudio 21, Naples 80125, Italy
[b] Institute of Science and Technology for Sustainable Energy and Mobility, National Research Council (CNR-STEMS), Piazzale Tecchio 80, Naples 80125, Italy
[c] Department of Chemical, Materials and Industrial Production Engineering (DICMAPI), University of Naples Federico II, Piazzale Tecchio 80, Naples 80125, Italy

ARTICLE INFO

Keywords:
Machine learning
Biomass pyrolysis
Generative adversarial network
Bio-oil
Missing data imputation
Explainable artificial intelligence

ABSTRACT

Biomass pyrolysis is a complex process, quite challenging to model physically and Modern AI methods could improve its prediction and characterization. However, AI model construction requires high-quality datasets. Existing datasets in literature, usually only a few hundred records, are inadequate for robust AI applications.

A first goal of the study was to make best use of the currently available body of experimental data on fixed bed non-catalytic biomass pyrolysis by comprehensively compiling available data from nearly 160 sources into a new dataset of 1137 records. Each record was carefully standardized to overcome inconsistencies in terminology and lack of uniformity among different sources. This extended dataset (including biomass properties, pyrolysis operating conditions, and bioliquid yield), integrating previous ones, is intended to promote community-based data sharing. The compiled dataset was characterized by remarkable data sparsity, due to lack of completeness of the original data.

A second goal was benchmarking different regression and data imputation models to assess the predictive ability of ML applied to the collected dataset. The most accurate estimates were obtained by leveraging a subset of about 500 instances without missing values, resulting in a Mean Absolute Error (MAE) of 2.28. Application of ML to the entire dataset with imputed missing data yielded a less accurate estimate (MAE = 3.45), a feature that underlines the criticality of missing data imputation, and of the sparsity of the dataset.

A third and mostly relevant goal was the critical assessment of Explainable Artificial Intelligence (XAI) techniques that come into play when ML is aimed at evaluating the importance and directional trends of selected features. XAI tools, namely Partial Dependence Plots (PDP) and SHAP, have been applied to the dataset to assess their trustworthiness to support mechanistic inference of the importance and directional trends of key biomass properties and process operational parameters on pyrolysis yields. The result of this analysis is far from satisfactory. Significant discrepancies across studies, inconsistencies among different methods and somewhat erratic trends in PDP plots reflect the challenge in achieving consistent mechanistic insights from purely data-driven approaches, suggesting the adoption of physics-informed machine learning embodying physico-chemical relationships to improved Explainable AI.

1. Introduction

In recent years, the energy sector has become increasingly digital and it is clear that further digitalization will be a key feature of the energy transition. Artificial intelligence (AI) is recognized as a key enabler to help the acceleration of the global energy transition. According to IEA report, in the coming years, AI will be decisive and will radically transform global energy systems, making them more interconnected, reliable and sustainable [1].

The development of renewable sources will help us cope with the current energy problems in the coming decades. However, large-scale applications can lead to uncertainties that threaten the reliability and stability of energy systems; in fact, renewable energy is characterized by strong volatility, intermittence and randomness. Therefore, forecasting renewable energies is essential to mitigate related uncertainties. Among the various forecasting techniques, the data-driven ones are attractive

* Corresponding authors.
E-mail addresses: antonioelia.pascarella@unina.it (A.E. Pascarella), antonio.coppola@stems.cnr.it (A. Coppola), stefano.marrone@unina.it (S. Marrone), roberto.chirone@unina.it (R. Chirone), carlo.sansone@unina.it (C. Sansone), piero.salatino@unina.it (P. Salatino).

https://doi.org/10.1016/j.fuel.2025.135000
Received 23 September 2024; Received in revised form 20 February 2025; Accepted 5 March 2025
Available online 18 March 2025

for their ability to shed light on hidden complex relationships among variables and for the ability to express them without using complex mathematics.

Artificial intelligence (AI) has recently triggered a paradigm shift in numerous sectors around the world. Machine Learning (ML), a branch of AI, is having a rapid spread in recent years, and represents a promising tool to manage some of uncertainties that characterize the renewable energy context. ML includes a set of techniques, mathematical models and algorithms able to extract information, identify patterns and predict potential outcomes of complex systems [2]. The use of ML techniques is being explored in the context of renewable energy for design, optimization, management, distribution and policymaking [3]. However, the use of these techniques in the energy sector is still limited to specific sectors. The literature suggests that most of the artificial intelligence developments for forecast application in the field of renewables are concentrated on the prediction of solar or wind energy [4]. Therefore, the true potential still remains to be expressed.

Among different kinds of renewable energies, bioenergy, abundant and sustainable, has attracted considerable focus [5]. Thermochemical conversion, such as pyrolysis, a process transforming biomass into biofuels cost-effectively and efficiently, has gained traction [6–8]. In pyrolysis, the feedstock is heated under an inert or oxygen-starving environment to produce bio-oil, biochar, and non-condensable gas. If upgraded appropriately, bio-oil could be an alternative to fossil fuels and a convenient source of platform chemicals. The diversity of biomass resources and the varying pyrolysis conditions dictate bio-oil properties [9], and modelling this process using ML can allow the prediction of bio-oil properties that derive from different biomass waste and can also enable plant optimization.

Artificial neural networks have been used to support pyrolysis modelling in the reduction of detailed or semi-lumped kinetic models, achieving a computational cost reduction by four orders of magnitude compared to the direct solution of the model, while keeping substantially unchanged the prediction accuracy [10]. This approach enables the integration of detailed kinetic models into broader and more comprehensive simulations, which also account for transport phenomena, thereby facilitating the optimization of pyrolysis processes at an industrial scale. Applications of pyrolysis related to sewage sludge are also highlighted, as in [11], where machine learning is employed alongside experimental data gathered from thermogravimetric analysis (TGA) to optimize process parameters, thus improving the yield of the desired products. Additionally, studies such as [12] illustrate the use of machine learning, appropriately integrated with physics-based simulation in Aspen Plus, for modeling other thermochemical processes, such as gasification.

Most of the relevant literature [13–16] on the application of ML techniques to the quantitative assessment of biomass pyrolysis focuses on predicting product yields, primarily bio-oil, to a lesser extent biochar, and less frequently bio-gas. A vast body of data refers to non-catalytic tests carried out in fixed bed pyrolyzers, the most diffused type, although an increasing number of observations refer to fluidized beds systems. Most of the tested biomass is of lignocellulosic nature. The input generally used for the prediction of the outputs are ultimate (UA) and proximate analysis (PA) coupled with pyrolysis conditions, such as pyrolysis temperature (PT), heating rate (HR) and average particle size of the biomass (PS). In a limited number of cases only, macro-components of biomass (Cellulose – Ce, Hemicellulose – He, and Lignin – Li) have been considered as input data, probably because quantitative assessment of biomass macro-components requires more demanding analysis than PA and UA. Many authors have employed classical ML techniques [17–21] for predicting bio-oil production, demonstrating its feasibility; however, some contrasting results have been observed. The analyses by different research groups regarding the significance of various inputs vary, especially concerning the relative importance of Proximate Analysis (PA) and Ultimate Analysis (UA).

This study aims at contributing to better application of ML tools to prediction and interpretation of biomass pyrolysis along three paths.

First, an extended dataset for fixed-bed pyrolysis of biomass is made publicly available after an extensive review of the scientific literature. The resulting dataset compiles 1137 records from 160 original sources. It is offered to the scientific community with the hope that future studies may benefit from a community-based approach to data sharing, so as to avoid having many small datasets scattered throughout literature. As also encouraged in [22], sharing quality data across diverse research communities is a prerequisite for practically implementing ML approaches in the biorefinery sector. During the data collection phase, inconsistencies in the terminology were identified in the literature. Moreover, incomplete and sparse experimental dataset may jeopardize effective application of ML methods, even when tools for missing data imputation are used.

Second, benchmarking of alternative prediction and missing data imputation methods has been accomplished with the aim of assessing ML predictive ability. The criticality of data sparsity and of the closely associated missing data imputation methods is recognized highlighting the urgent need for more consistent practices and standardized approaches in reporting experimental data.

Third, the potential of ML methods to support interpretative analysis of pyrolysis data, by characterization of the importance and directional trends of selected features, has been assessed. A critical comparison of studies based on application of Explainable AI methods to pyrolysis highlights discrepancies and inconsistencies that shed doubt, at present, on the trustworthiness of Explainable Artificial Intelligence (XAI) methods for interpretative purposes. Open challenges and research priorities related to AI applications in Pyrolysis are also presented and discussed.

2. Material and methods

As highlighted above, it was important to create a dataset in the field of biomass pyrolysis in order to unify the different datasets in the literature and boost the AI sector for bio-energy applications by creating a more comprehensive dataset. The following section presents the dataset and shows how the missing data was handled on the dataset and the establishment of a machine-learning benchmark on bio-liquid yield. Section 2.1 presents the dataset and proposes data collection guidelines to develop a common framework. Section 2.2 describes the missing data imputation frameworks, while Section 2.3 describes the models used for regression on the bio-liquid yield. Finally, Section 2.4 outlines the experimental setup.

2.1. Data collection

The dataset,[1] named Pyris, used in this work was compiled from experimental data obtained from about 160 research articles available in literature regarding non-catalytic biomass pyrolysis in fixed bed reactors, using nitrogen as sweep gas, for a total of 1137 observations. Specifically, the dataset was a combination of 4 different datasets already available in the literature from different research groups [19,21,23,24] and a further extension carried out by the authors. Any observation is labelled by type of the tested residual biomass; most of the biomasses present in the dataset belong to the Plantae kingdom (99.6 %) except for only 4 observations belonged to Bacteria and Chromista kingdoms. Table B.5 in the appendix shows a classification of the biomass for family from a taxonomy point of view, specifying the relative abundance and the different biomass type. Lacustrine alga, Spirulina Sp. (Phormidiaceae family) and Nannochloropsis (Eustigmataceae family) are the only types of biomasses not belonging to plantae kingdom. Asteraceae, Betulaceae, Poaceae, Oleaceae, Brassicaceae, and

[1] The dataset is available to reviewers upon request and will be made public after acceptance.

Fabaceae are the families mainly represented in the dataset (56.2 %) which include biomass type such as safflower seed, sunflower bagasse, hazelnut shells, olive residues etc.

For each observation 18 different features have been collected which can be classified in: 1. biomass properties; 2. pyrolysis conditions; and 3. pyrolysis performances. Biomass properties in turn can be categorized in Proximate analysis (Ash, Fixed Carbon and Volatile matter on dry basis),[2] Ultimate analysis (Carbon − C, Hydrogen − H, Nitrogen − N, and Oxygen − O calculated by difference), and macro-components (cellulose − Ce, hemicellulose − He, and lignin − Li) typical of ligno-cellulosic biomass have been collected, all expressed as percentages on a mass basis.

Pyrolysis temperature (°C) and Heating rate (°C/min) have been considered for understanding the effect of pyrolysis conditions on process performance. Furthermore, the biomass particle size was classified as pyrolysis conditions rather than being considered as an intrinsic property of biomass. This choice was made since it has significant effects on the biomass heating rate and hence on the overall pyrolysis performance.

A separate discussion must be made for the Sweep gas flow rate. it is remarkable that some previous studies included this variable in their analysis. It is likely that the sweep gas flow rate affects bioliquids yield via its effect on concentration and residence time of pyrolytic vapours in the pyrolytic converter. To properly assess the influence of this variable, which is inherently extensive − i.e. dependent on the scale of the apparatus − the sweep gas flow rate should be scaled by reference to variables expressing the reactor size or capacity. This has not been done in previous studies, raising some concern as to the reliability of the predictions. The impact of the sweep gas flow rate was not considered in the present study as it was not possible to define a meaningful scaling parameter that could be applied to both continuous and discontinuous operation of pyrolytic converters. This point is worth of consideration in future studies.

The quantitative distinction between the organic phase and the aqueous phase requires a separation process, for example by extraction or centrifugation, which not all experimental campaigns carry out. Consequently, the authors decided to compile the dataset reporting the value of bio-liquid as pyrolysis yield, and where a distinction between the organic and aqueous phase is present, the authors combined the data to obtain the total bio-liquid. This choice allowed to have a higher number of observations available.

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. Obviously, also other properties are crucial to understand the possible utilization and/or upgrading of the bio-liquid, such as organic phase compositions, pH, viscosity, pour point and flash point etc., however these latter are reported only very occasionally, and sometimes measured under different conditions. For these reasons, the hydrogen and oxygen content of the organic phase, although they have been collected and present in the dataset, they are not considered in this work.

With bio-liquid yield, the other output investigated in this work was the aqueous phase yield expressed as weight percentage of the aqueous phase respect to the initial biomass, albeit presenting a number of observations approximately one third of that relating to bio-liquid (413 observations). Table B.6 in the appendix shows a sample of the dataset described above. The relative abundance of the different features in the dataset is reported in the Tab.B.7 in the appendix.

Concerning the biomass properties, the Ultimate analysis is practically always present with a few exceptions, the Proximate analysis shows an abundance of about 88 % on the total of observations, while macro-components have an average abundance of about 65 %. Conversely, the 4 features used for the characterization of the pyrolysis process the abundance is 100 %. The management of the missing data are discussed in the following paragraph. In the end, Table 1 and Fig. 1 report the statistical characteristics of the collected data, including mean, standard deviation, minimum and maximum values, quartiles, and correlations, which will be useful for the discussion of results and experimental setup. Finally, considering the intrinsic correlation that exists among some variables as they are calculated as a difference from the others (such as for example the data relating to the proximate and ultimate analysis), fixed carbon, oxygen and hemicellulose were preventively excluded from the dataset used for the training.

2.2. Missing data imputation

The current study collected data from the literature, and various public datasets were integrated into a unified dataset. Unfortunately, due to incompleteness of several records, the compiled dataset resulted in a fairly sparse matrix. This matrix needed to be managed to construct a machine learning system to predict bio-liquid yield using biomass properties and operating conditions.

The missing data problem was addressed in addition to basic methods as imputing with mean by employing two distinct strategies: imputing with Generative Adversarial Networks [25] or using an iterative method in a round-robin fashion. GAIN approaches the missing data imputation problem by leveraging the Generative Adversarial Networks (GANs) framework [26]. The method utilizes two neural networks: the generator and the discriminator. The generator aims to impute missing values, while the discriminator attempts to distinguish between imputed and observed elements. These two networks are trained in an adversarial game where the generator learns to produce increasingly plausible data to fool the discriminator, and the discriminator becomes better at detecting imputations. The trained generator is used to fill in missing values. The iterative method in a round-robin fashion, as explained in the scikit-learn iterative-imputer doc,[3] is a framework inspired on [27], that is described in the following. The term'round-robin' describes a situation where resources are allocated in an equal, cyclic manner; in this case, each variable is addressed in turn, with the missing values being filled using the remaining variables as predictors.

Each missing value was initially provisionally populated using basic estimates such as the variable's mean, median, or mode. Subsequently, a variable with missing values (for this explanation, this will be referred to as Variable A) was selected, and its preliminarily filled values were treated as missing once again. The other variables were then used to predict the missing values for Variable A, including those that initially had missing values but were replaced with estimates in the first step. This process is repeated for all variables until the convergence criteria are met. In this study, this step involved the use of two imputation models using the imputation strategy described above: Random Forest [28] (also known as *MissForest* when used to impute missing data) and K-Nearest Neighbors (KNN) [29]. Random Forest is also used for the regression task, which will be better described in the subsequent section; the K-Nearest Neighbors (KNN) algorithm was used only in the data imputation framework to impute missing values, and it is based on the mean of the nearest K points.

2.3. Regression models

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 a concise overview is provided. Random forests (RF) offer an enhancement over bagged trees by introducing a minor modification that reduces the

[2] All the pyrolysis tests in the dataset were carried out on previously dried biomass.

[3] https://scikit-learn.org/stable/modules/impute.html#iterative-imputer.

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

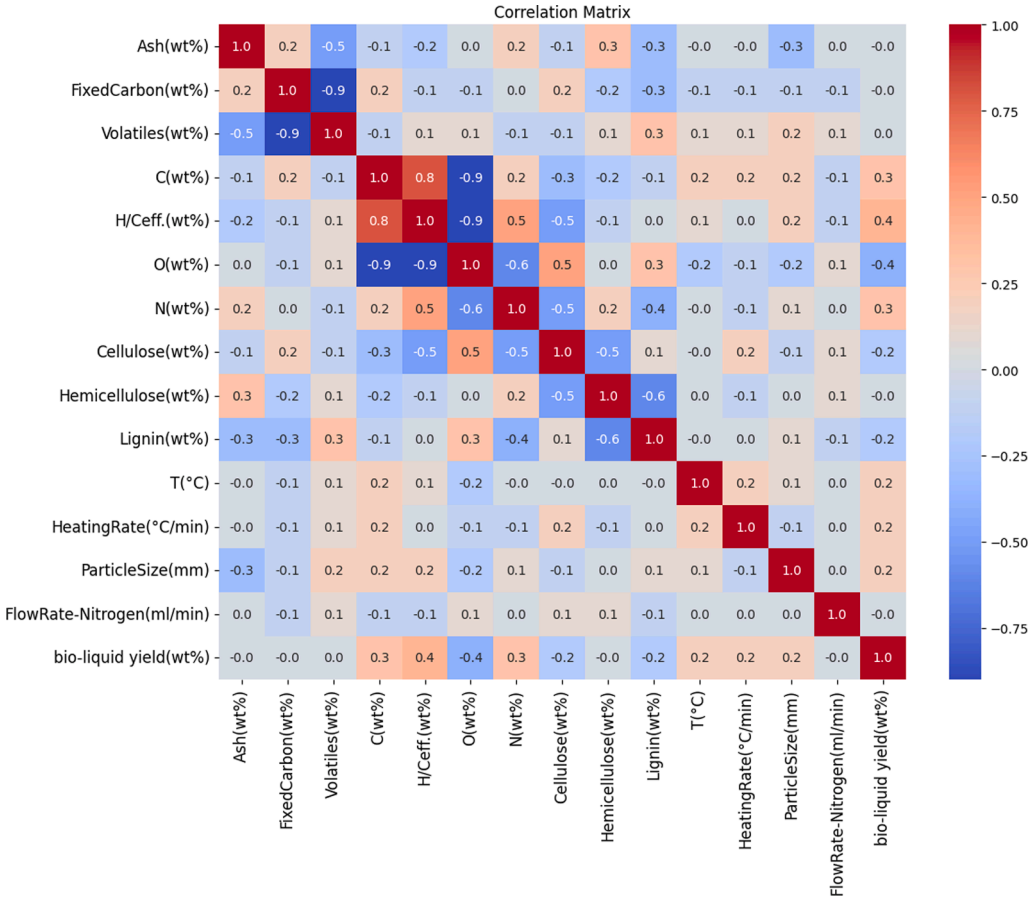


Fig. 1. Correlation matrix of the data collected.

correlation among the trees. Bagging is the foundation for random forests, an ensemble technique where models are trained on various datasets created through bootstrapping (re-sampling with replacement) and aggregated. In Random Forest, each time a split in a tree is deliberated, a random selection of m predictors is drawn from the entire pool of p predictors to serve as split candidates. Typically, m is chosen to be approximately the square root of p [30]. While Support Vector Machines are traditionally recognized for classification, they can be effectively adjusted to handle regression tasks and, in this case, are referred to as Support Vector Regressor (SVR). Part of the method involves an 'epsilon-insensitive' error measure. The error measure dismisses errors below a certain threshold, called epsilon. This characteristic means that only errors of significant size are considered when adjusting the model. Such an approach can be visualized by imagining a region around a line where any data point inside the area is not considered an error [31]. In SVR, two key hyperparameters are crucial: epsilon, which defines the error insensitivity threshold, and a regularization parameter that balances model fit and complexity to prevent overfitting. Boosting is a

robust machine-learning framework primarily used in supervised learning. It functions by aggregating multiple weak predictors to form a strong predictor. These weak predictors are typically decision trees. The algorithm learns iteratively from the mistakes of the previous predictors and adjusts the predictions based on these errors. The process continues for a predefined number of iterations. The final model is the weighted sum of all the predictors [30,32]. The perceptron model represents the foundational structure in neural networks. At its simplest, it consists of an input layer that elaborates and spreads the input information to an output layer. However, the multilayer perceptron (MLP) becomes essential for more complex applications. Multilayer neural networks are constructed with several sub-layers, where each node in one layer is connected with every node in the subsequent layer. Instead of relying solely on linear functions, MLP can utilize diverse functions like the sigmoid, hyperbolic tangents or relu in its hidden layers. The training of MLP employs backpropagation. This method processes inputs in the forward phase to produce a predicted output and then compares it to the actual label. In the backward step, errors from this comparison lead to weight adjustments, enhancing the neural's training [33].

The Mixture of Experts (MoE) model [34] in machine learning is an ensemble approach where multiple specialized models, or'experts,' tackle different parts of a complex problem. A key component is the gating network, which determines which expert handles which data instance based on its characteristics. This approach allows each expert to focus on a specific subset of the task. MoE models are beneficial in natural language processing, where they can adaptively address various aspects of language but can also be applied to tabular data.

2.4. Experimental setup

To establish model benchmarks on the regression of bio-liquid yield on the novel dataset introduced, the experiments were conducted in both scenarios of the original dataset with imputed missing values and with a subsample of all rows without missing values, resulting in approximately 500 instances, to highlight the performance differences with or without the noise induced by data missingness.

From the original dataset, three imputed dataset versions were produced; the imputation was performed with GAIN, *MissForest* and KNN. Iterative Imputer from scikit-learn was used to implement data imputation with *MissForest* and KNN, while PyTorch was used to implement GAIN. XGBoost, Random Forest, Support Vector Regressor, Multilayer Perceptron, and a Mixture of Experts were utilized to achieve bio-liquid yield regression on the imputed data and were implemented respectively using the libraries xgboost for XGB, scikit-learn for RF, SVR, and MLP and PyTorch for MoE. Variables with a Pearson correlation coefficient less than 0.3 in absolute value were considered to have negligible correlations [35]. Choosing non correlated features is necessary for a proper interpretability analysis conducted in the below sections. The selection of variables was made considering the correlation matrix of the dataset without missing to make a robust choice concerning the problem of missing values.

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 [24], genetic algorithms were used to optimize hyperparameters and not for feature selection. 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.

The search for genetic algorithms was constructed to maximize a fitness function, equal to minus the mean absolute error among predicted and actual values, calculated for each cross-validation iteration on the validation set, obtained in each cross-validation step by further dividing the training into training and validation with an 80–20 ratio. The genetic algorithm was implemented using the "ContinuousGenAlgSolver" class from the general library. The parameters used for the genetic algorithm were set equal to the default ones for mutation rate, selection rate, and selection strategy. Meanwhile, the population size was 25 instead of 10 (default) for a broader spectrum of possible hyperparameters. The maximum generations were set to 10 instead of 200 both to avoid overfitting on the validation set and because, after ten generations, the research procedure started to converge; the number of genes was set equal to the number of hyperparameters to optimize. The learning rate and regularization parameters were initialized on a logarithmic scale base 10. The code can be accessed at the following link[4].

3. Results and discussion

As explained, it was crucial to gather various datasets found in the literature, resolve discrepancies, consolidate them into a single dataset with additional instances, and share it with the ambition of fostering a community-based approach among researchers who may be interested in implementing the dataset by sharing additional data. Section 3.1 offers a comprehensive dataset benchmark, elucidating each model's performance metrics for predicting bio-liquid yield. Section 3.2 discusses the significance of the features and elucidates the trends of bio-liquid yield as a function of these individual variables. Section 3.3 discusses some critical aspects of AI applications in Pyrolysis and the open challenges defined on the shared dataset to encourage the development of machine and deep learning algorithms tailored for the Pyrolysis domain.

3.1. Model benchmarks

Comparisons were made among XGB, MLP, MoE, RF, and SVR on the original data where missing values were filled with various missing-data imputation frameworks and with a reduced dataset with only instances without missing values, resulting in a dataset of about 500 samples. The study was conducted on these two scenarios to understand how missing values impact the models' performance and the downstream interpretative analysis. It was observed that XGB model, with GAIN as a missing data imputation framework, emerged as the most proficient for estimating bio-liquid yield, as displayed in Table 2 in the context of the original dataset. XGB performed better with the reduced dataset, as shown in Table 3.

In Tables 2 and 3 are shown as best performances, a mean squared error, root mean squared error, mean absolute error and r-squared respectively of 32.17 $\pm$ 4.21, 5.66 $\pm$ 0.38, 3.45 $\pm$ 0.34 and 0.66 $\pm$ 0.04 for the bio-liquid yield for the original dataset and 17.80 $\pm$ 11.54, 4.21 $\pm$ 1.21, 2.28 $\pm$ 0.58 and 0.80 $\pm$ 0.12 for the bio-liquid yield for the reduced dataset.

Fig. 2 reports the parity plots of predicted values of the bio-liquid yield versus actual observed values in these two scenarios; in both cases the regression model employed was XGB. It should be noted that in this context of tabular data with about one thousand samples, although not in a proper Big Data context, the most successful methodologies have

[4] https://github.com/priamus-lab/PYRIS.

Table 2
bio-liquid yield benchmark on the original dataset.

imputation model	prediction model	Mean squared error	Root mean squared error	Mean absolute error	R^2
GAIN	MoE	170.83 ± 17.54	13.07 ± 0.65	11.54 ± 0.68	0.00 ± 0.27
	XGB	**32.17 ± 4.21**	**5.66 ± 0.38**	3.66 ± 0.33	**0.66 ± 0.04**
	RF	33.82 ± 4.57	5.80 ± 0.39	3.97 ± 0.25	0.65 ± 0.05
	SVR	53.50 ± 12.32	7.26 ± 0.88	4.45 ± 0.43	0.45 ± 0.10
	MLP	51.20 ± 8.19	7.13 ± 0.57	5.10 ± 0.50	0.47 ± 0.06
Miss Forest	MoE	172.49 ± 31.37	13.13 ± 1.15	10.56 ± 0.91	0.00 ± 0.32
	XGB	43.81 ± 10.67	6.57 ± 0.83	**3.45 ± 0.34**	0.54 ± 0.10
	RF	34.22 ± 5.77	5.83 ± 0.50	3.73 ± 0.36	0.64 ± 0.06
	SVR	68.03 ± 8.88	8.23 ± 0.54	5.06 ± 0.39	0.29 ± 0.09
	MLP	56.16 ± 6.06	7.48 ± 0.40	5.33 ± 0.28	0.41 ± 0.07
KNN	MoE	164.87 ± 20.52	12.84 ± 0.78	10.36 ± 0.79	0.00 ± 0.35
	XGB	54.94 ± 9.51	7.39 ± 0.63	3.75 ± 0.35	0.42 ± 0.12
	RF	40.88 ± 5.04	6.38 ± 0.39	4.05 ± 0.35	0.57 ± 0.06
	SVR	71.28 ± 7.22	8.43 ± 0.43	5.21 ± 0.20	0.25 ± 0.14
	MLP	64.28 ± 6.23	8.00 ± 0.39	5.82 ± 0.30	0.32 ± 0.10
Mean	MoE	167.12 ± 20.74	12.92 ± 0.79	10.53 ± 0.85	0.00 ± 0.32
	XGB	40.78 ± 9.11	6.34 ± 0.71	3.58 ± 0.33	0.57 ± 0.09
	RF	34.30 ± 5.60	5.84 ± 0.47	3.71 ± 0.37	0.64 ± 0.06
	SVR	60.10 ± 9.36	7.73 ± 0.59	4.76 ± 0.44	0.37 ± 0.09
	MLP	56.65 ± 16.80	7.46 ± 1.02	5.22 ± 0.46	0.39 ± 0.21

Table 3
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

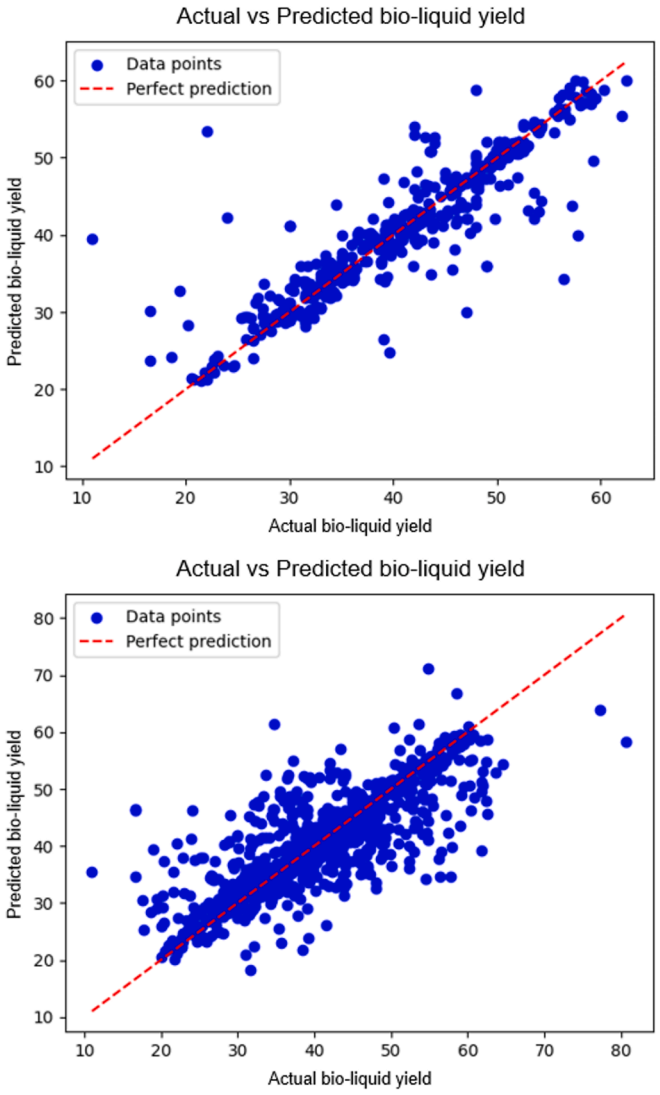


Fig. 2. Predicted versus observed bio-liquid yield. Top) reduced dataset without missing data (468 records); Bottom) original dataset (1137 records).

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.

As can be seen from Tables 2 and 3, the performance metrics in the case of the experiments on the reduced dataset without missing values improve, and this is linked to the fact that noise in the data is reduced; in particular, this aspect will be emphasised even more in the interpretability analysis, as shown in the following section, highlighting the fact that more robust handling of missing data to reduce noise in the dataset becomes a critical task.

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.

3.2. Effect of input variables on bio-liquid yield, using Explainable Artificial intelligence (XAI)

Further to the development of a system capable of predicting bio-liquid yields from waste biomass, it would be desirable to open up these black-box systems to understand how the system produces outputs as a function of inputs, giving more reliability to the predictive system by showing whether its outcomes are compatible with domain knowledge by leveraging methods from Explainable Artificial Intelligence (XAI). Partial Dependence Plots (PDP) are the most recurrently used interpretability methods in scientific literature [21,36–38] on machine learning applications to pyrolysis. If features of a machine learning model are correlated, the partial dependence plot cannot be trusted. The computation of a partial dependence plot for a feature strongly correlated with other features involves averaging predictions of artificial data instances that are unlikely in reality with the possibility to bias the estimated feature effect [39]. Accordingly, 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 [35]. Ash, H/C_{eff},[5] Cellulose, Lignin, Temperature, HeatingRate, ParticleSize were selected for this analysis. The PDP plots for the different features that predict the bio-liquid yield are reported in Fig. 3. The dataset used for the interpretative analysis is the reduced version without missing values, resulting in approximately 500 samples, both because, on this dataset, the metric performances of prediction were better as shown in Table 2 and 3 and because, as Fig. A.6 in the appendix also shows, the PDP produced on the full dataset imputed with GAIN sometimes produced results that differed from the domain knowledge, due to the noise introduced by the presence of missing data. The discussion of the PDP plots in Fig. 3 is reported in the section below in a schematic manner and individually for each input variable, making specific reference to the behaviour shown in the range between fifth and ninety-fifth percentiles (see Table 1). The reference percentile values are reported in brackets next to each variable for convenience. The results obtained in this study are compared with similar findings from other research groups, as summarized in Table 4, and appropriately discussed according to the pyrolysis mechanism.

The MAE calculated with ten-fold cross-validation was assumed as reference error, equal to about 2 % of the bio-liquid yield. Below this threshold, any variation in the trend is considered insignificant because it is below the inherent model's noise.

Feature importance was reproduced using XGBoost and SHAP (SHapley Additive exPlanations) [41], as illustrated in Fig. 4. The feature importance for XGBoost was derived using the XGBoost library in Python, utilizing the'importance type' parameter set to'weight', which counts the number of times features appear across the trees. On the other hand, SHAP provides a more robust approach to feature importance assessment [42]. It relies on the Shapley values from game theory, ensuring a fair distribution of importance across features. A critical mathematical property that sets SHAP apart is its consistency. If a model changes in a way that makes a feature more critical, the SHAP value for that feature will not decrease [43]. This property makes SHAP reliable for feature importance analysis compared to traditional methods like those used by XGBoost. The SHAP values were generated using the SHAP library. Table 4 suggests that many studies addressing machine learning applied to biomass pyrolysis utilized the importance of the Random Forest feature to determine the significance of variables. However, in the light of the consistency property explained in [43], and considering the contradictions that emerge when comparing Random Forest with SHAP, see Fig. 4, together with the consideration that SHAP is a more robust theoretical framework [43], it may be advisable to adopt SHAP feature importance. The use of non-robust interpretability methods, as well as the use of small datasets, could explain the recurrent discrepancies in establishing the importance of features, as shown in Table 4.

3.2.1. Biomass properties

Ash: The PDP plot reveals a negative effect on bio-liquid yield with increasing ash content in the biomass. This was also observed by others [20,44], and is consistent with the role of ash as catalyst promoting secondary cracking of pyrolysis vapors reducing liquid yield as well as liquid quality [45]. A slightly positive effect may be observed for ash content lower than 6 %, although variations are within a 2 % error range. A similar trend has also been identified by other research groups [19,24]. Fig. 4 shows that the effect of ash on the bio-liquid yield is important for both methods. This agrees with the results of some authors [19,20,44] and disagrees with others [21,38,46]. *H/C_{eff}*: As documented in the literature [47–50], the increase in this parameter appears to have a positive effect on the production of bio-liquids, conversely limiting the production of coke. The PDP graph confirms that the trained model is able to predict this behavior, with the exception of a first point which however falls within the error of the model. The importance of this parameter is also confirmed by the analysis of the future importance with both the XGboost and SHAP methods (see Fig. 4). Direct comparison of this feature with other authors is not possible because the H and C values are generally used separately for training. Carbon appears to have a low impact for some research groups [19,20,46], while for others it has a medium–high impact [21,24,38,44]. Similarly, for hydrogen, different relevance is found by the different research groups.

Cellulose and Lignin: These two features have 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, where a crucial role is played by the reaction conditions [51,52]. The 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 [53,54] with fairly small H/C_{eff} ratio [55] that may favour the production of further bio-char at the expense of bio-liquid.

3.2.2. Pyrolysis conditions

Temperature, T: The effect of temperature on the yield of pyrolysis products is among the most studied in the literature, with observations typically converging toward an indication of 500–550 °C as the temperature range within which bio-liquid yield is maximized. The PDP graph is fully consistent with this observation. The SHAP method ranks the importance of temperature at a medium level, while XGboost suggests it to play a more important role.

Heating rate: Like T, the heating rate is considered one of the key operational variables, much like a "knob" to switch from slow to fast pyrolysis, which involves a greater yield of bio-liquid as its value increases. This is consistent with indications of the model reported in the PDP graph. The SHAP method indicates that the heating rate is of medium importance, whereas XGBoost ranks it among the least influential variables. A certain spread in the importance of this variable is reflected also by previous studies, that attribute high importance [20,21,46], moderate importance [19,38] with only a few attributing low importance to this parameter [44].

Particle size: According to the SHAP method, this variable has a low impact, with its variability well within the range of the error of the prediction model. The effect of particle size reported by other research groups is rather erratic, with some studies indicating a non-monotonic trend [19–21,24].

[5] hydrogen/carbon effective ratio: the choice to use this variable depends on the fact that it appears to play an important role in converting biomass to liquid fuels efficient [40].

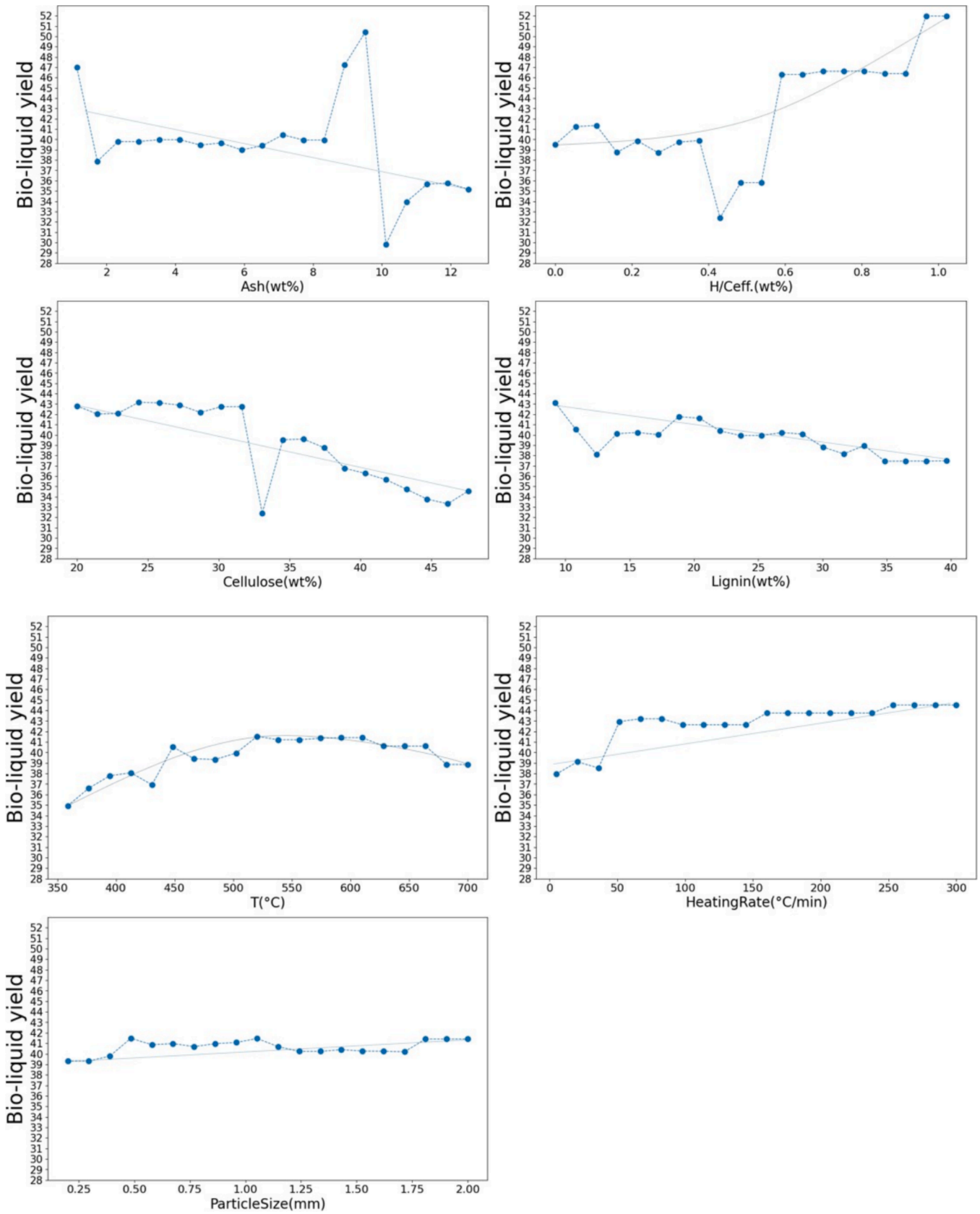


Fig. 3. Partial dependency plots (PDP) for bio-liquid yield developed on the reduced dataset without missing data with respect to biomass properties and operating conditions. The solid line is only a guide-to-eye.

Table 4
Comparison of feature importance and effects on bio-liquid yield across studies. Feature importance is shown as [position]/[total variables] (e.g., 3/8 is third of 8). Effects on yield: Negative (↘), Positive (↗), non-monotonic trend with a minimum (↘↗), with a maximum (↗↘), Not reported (NR). Symbols: + analysis based on H/C ratio; * distinguishes slow and fast pyrolysis; U for ultimate analysis; P for proximate analysis; ** indicates analysis on reduced dataset without missing values; *** analysis based on H/Ceff. ratio. SHAP, RF F.I., and PDP refer to feature importance methods: SHAP values, Random Forest importance, and Partial Dependence Plots.

	Numberof observations	Ash	Cellulose	Lignin	Hydrogen	T	Heating rate	Particle Size	Flow rate nitrogen	XAI
This work	468 (reduced)** 1137 (original)	3/7 ↘	1/7 ↘	6/7 ↘	2/7*** ↗	5/7 ↗↘	4/7 ↗	7/7Fluctuations in error margins	NR	SHAPPDP
[46]	122	5/8	2/8 ↗	7/8	NR	1/8 ↗↘	NR	8/8	6/8 ↗	RFF.I. PDP
[21]	282	NR	NR	NR	1/7 ↘↗	7/7 ↗↘	2/7 ↗	3/7 ↘↗	NR	RFF.I. PDP
[20]	292	3/6 ↘	NR	NR	NR	6/6 ↗↘	1/6 ↗	4/6 ↘↗	5/6 ↗	SHAPPDP
[44]*	217	3/13 ↘	NR	NR	7/13 +	2/13 ↗↘	NR	NR	NR	SHAPPDP
[38]	322	NR	NR	NR	NR	1/8 ↗↘	3/8 ↗↘	NR	NR	PDP
[24] U	263	NR	NR	NR	↘↗	↗↘	↗	↗	↗	PDP
[24] P	263	↗↘	NR	NR	↘↗	↗↘	↗	↗	↗	PDP
[19]	264	3/7 ↗	NR	NR	NR	7/7	4/7 ↗	5/7 ↘↗	6/7 ↗↘	RFF.I. PDP

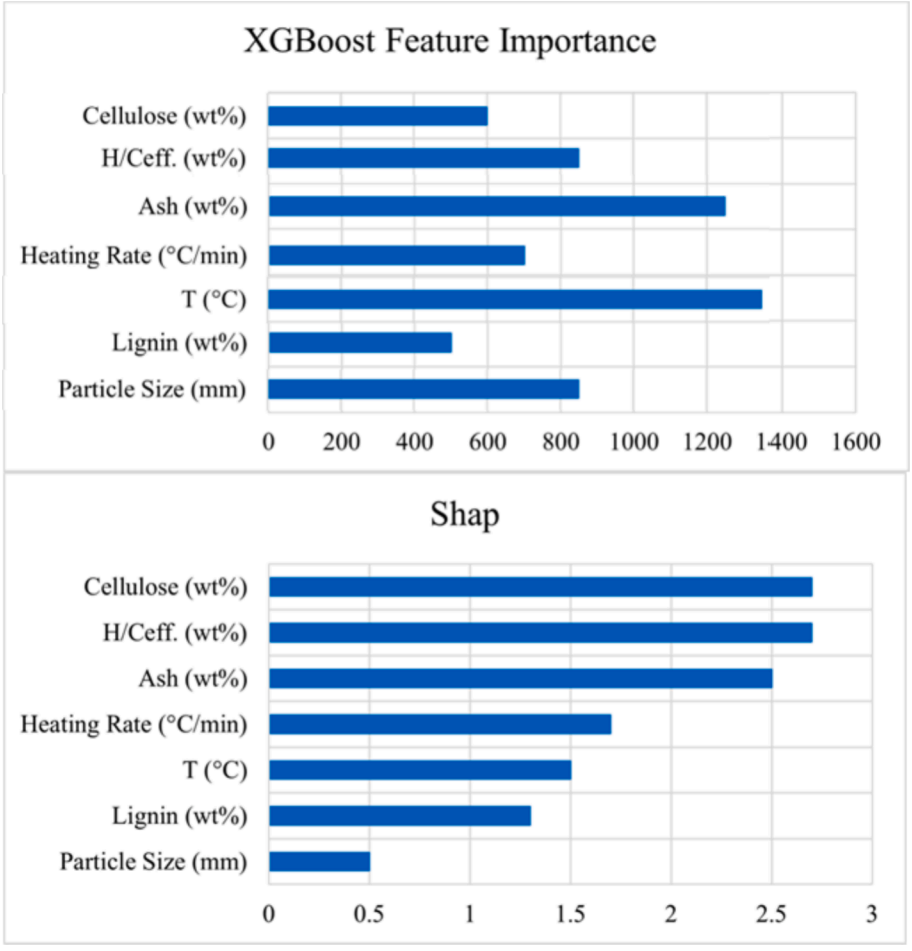


Fig. 4. Comparison of variable importance methods: the importance of variables for predicting bio-liquid yield is obtained with XGB; feature importance is computed using SHAP. This comparison highlights the contradictions between two methods of XAI, emphasizing the robustness of SHAP.

3.3. Critical discussion on application of AI to pyrolysis data

There is considerable effort to integrate AI into scientific research. AI can be a powerful tool for generating predictions in scientific and engineering fields. However, caution must be exercised to avoid using it naively, as uncontrolled application of AI in the fields of natural and engineering science can raise several threats [56]. There is evidence of considerable over-optimism in scientific claims based on machine

learning model performance [57], likely due to a poor understanding of the limits of machine prediction in fields outside computer science. Hog and Villars [58] express the warning that, from an epistemological perspective, AI is judged to be performant solely based on its metrics on the data. At the same time, in the fundamental sciences, a model is considered good not only if it agrees with the experimental data, but also if its formal structure is sound and if it may be well integrated with other theories. They recognize that machine learning has developed significantly within industrial and commercial contexts where "forecasting" is prioritized over "understanding", and emphasize that a crucial effort in machine learning for natural and engineering sciences should regard making it more "physically aware" to enhance its reliability. It has been argued [59] that the acceptance of AI models in the context of process engineering is contingent upon their alignment with domain knowledge. When restricting these general warnings to the case at hand, namely pyrolysis of biomass, results in sections 3.2 indicate that the use of commonly available Explainable AI methods does not bring new and significant insights for biomass pyrolysis, at least at their present stage of development. Introducing more "Physics-Awareness" in the AI pipeline might improve data interpretability by XAI. No step has been made so far, in this as well as in previous studies [17,19–21,24,38,44] along this track. This will certainly be a priority in future studies.

Moreover, critical issues were found in the present study during data collection, due to lack of uniformity and/or standardization. The very liquid product of pyrolysis lacks an unambiguous definition, as terms like bio-oil and bio-liquid are often used indiscriminately. In some studies, bio-oil and bio-liquid are used interchangeably to denote the whole stream of condensable products issuing from the pyrolyzer. In other cases, bio-oil refers to the hydrophobic organic phase, whereas bio-liquid refers to all the condensable matter, which embodies also the aqueous fraction consisting of water (initial moisture plus pyrolysis water) and the hydrophylic water-miscible fraction of the organic compounds. In compiling the data, the authors followed the latter definition, by assuming bio-liquids as the total condensed products and bio-oil as the hydrophobic water-immiscible fraction of the bio-liquid. Retrieval of data from the open literature is also frequently hampered by the circumstance that data are presented exclusively in graphic rather than tabulated form, which makes their reading not immediate (indeed in many cases specific software was used to the extraction of graphic data). More generally, lack of homogeneity in the data entails an intrinsic difficulty in collecting them, impacting both on the quality of the final dataset which may be affected by errors, and on the difficulty of implementing systems for automatic data collection. Data collected from various laboratories might not uniformly report all measurements, especially regarding biomass properties, hence, the dataset exhibits some sparsity. The problem of handling missing data, as demonstrated in this study, is particularly crucial because the noise introduced by the presence of data sparsity not only deteriorates the performance of the models but also jeopardizes the interpretability analysis. Therefore, considering that in AI literature, when sharing datasets with the scientific community, the related tasks are also defined [60], the open challenges related to this dataset are better discussed here. In the light of the above arguments, it would be highly advisable, therefore, to have some sort of agreement of the biomass pyrolysis scientific community, as in other areas, in terms of common definitions and usability of data, so as to encourage the growth of a common dataset instead of many small independent ones. This is also one of the objectives of this work with the sharing of the PYRIS dataset.

No doubt, the most challenging goal is pursuing AI beyond forecasting purposes as a tool to clarify relationships among variables and to support mechanistic inference. This goal is very ambitious and, as already discussed, still problematic. Directional trends in one-way partial dependence between variables and scoring of the importance of parameters vary significantly from study to study. One of the main reasons is the use of small and scattered datasets, a limitation that we tried to overcome by sharing with the community a more extended dataset encompassing various datasets present in the literature. Moreover, it must be considered that purely data-driven approaches can produce physically inconsistent solutions [61]. Table 4 demonstrates how different machine learning solutions have proposed trends of bio-liquid yields relative to input variables in a inconsistent manner. This risk must be emphasized and calls for additional research and development of novel algorithms to produce physically informed solutions. This is being done in the context of AI applications for biomass gasification, where a physics-informed neural network method (PINN) has been developed to predict biomass gasification products by embodying physical constraints in the loss function, providing physically feasible predictions in [62,63].

4. Conclusions

The application of machine learning to the prediction of yield and quality of products from biomass pyrolysis has been recently demonstrated. However, issues have arisen due to the lack of a large reference dataset, with only many small datasets scattered throughout the literature. Moreover, various studies show discrepancies regarding the different trends that models reproduce regarding bio-liquid yield relative to biomass properties and process operational conditions. The present study tries to overcome the shortcomings of using small datasets by reporting an extended dataset of 1137 independent records based on experiments of non-catalytic fixed bed pyrolysis of biomass (mostly of lignocellulosic nature). The dataset includes the properties of the biomass (proximate and ultimate analysis), pyrolysis conditions (pyrolysis temperature, heating rate, particle size), and yield of the bio-liquid (assumed as the sum of the organic and aqueous phase). The dataset, the largest of its kind, embodying previously published ones, is shared with the scientific community as a basis for further studies, thereby encouraging a community-based approach.

The study highlights some critical issues associated with data retrieval from published literature. There is a certain degree of heterogeneity in the definition of the variables, such as the lack of a standard definition of the liquid products of the pyrolysis process, and discrepancies in the presentation of the data, either in tabular or graphical form. Moreover, published studies might not uniformly report all measurements, especially regarding biomass properties, giving rise to sparse datasets. Although missing data algorithms may be used to manage sparse information, this may only happen at the expense of accuracy of the models, especially as far as Explainable Artificial Intelligence is the aim of the study. Standardization in the definition of variables, better uniformity in the presentation of results, completeness of data are critical prerequisites for application of AI-based tools for automatic data retrieval from scientific articles.

The present study experiments were conducted in two scenarios: the original dataset with missing data imputed and a subset without missing data. In the first scenario, the overall best performance was achieved using the imputed dataset with GAIN and the XGBoost regression model, resulting in a mean absolute error (MAE) of 3.66 and an R-squared of 0.66. In the second scenario, the best regression model was again XGBoost, yielding a MAE of 2.28 and an R-squared of 0.80. The analysis of the PDP plots highlighted better congruence between the influence of the different variables on the investigated bio-liquid yield and chemical-physical mechanisms relevant to pyrolysis in the scenario of the data subset without missing data. In contrast, inconsistencies emerged during model interpretability analysis of the original dataset with imputed data, caused by the missing data noise and documented in the Appendix.

Comparison of results generated by the various research groups, reported in Table 4, highlight remarkable discrepancies in feature importance and directional trends. The source of these discrepancies has been discussed, possibly related to the characteristics of the different datasets, in terms of size and distribution of the data, and to the use of less robust XAI methods, such as the feature importance of Random Forest instead of SHAP [43]. Closely related to the previous issue, are the

discrepancies between the different studies in the way machine learning models represent the relationships between variables through one-way partial dependence analysis. This criticality reflects a more general caution about using purely data-driven machine learning 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.

CRediT authorship contribution statement

Antonio Elia Pascarella: Writing – original draft, Methodology, Formal analysis, Data curation. **Antonio Coppola:** Writing – original draft, Methodology, Formal analysis, Data curation. **Stefano Marrone:** Supervision. **Roberto Chirone:** Supervision. **Carlo Sansone:** Supervision. **Piero Salatino:** Writing – review & editing, Supervision.

Declaration of competing interest

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.

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'Università e della Ricerca (MUR), CUP - B53C22004060006, CUP - E63C22002160007, Project title "Network 4 Energy Sustainable Transition – NEST". Antonio Elia Pascarella and Roberto Chirone acknowledge support from Ministero dell'Università e della Ricerca (MUR), in the frame of PON "Ricerca e Innovazione" 2014-2020, Actions IV.5 and IV.6. The research activity was performed in cooperation with Eni S.p.A. in the frame of the Agreement 4400007890 between Eni S.p.A. and Università degli Studi di Napoli Federico II.

Appendix A

Fig. A.5 reports, for completeness, the dataset's correlation matrix without missing data utilized for interpretability analyses. Fig. A.6 shows the PDP plots obtained using the entire dataset, whose missing data were imputed using GAIN (Generative Adversarial Imputation Nets). The presence of missing data certainly alters the trend predicted by the PDP plots of the selected variables compared to the case in the total absence of missing data, particularly for ash, lignin, and cellulose. These variables present a non-negligible number of missing data equal to 12, 36, and 23 %, respectively, which almost certainly influences their influence in the model. The remaining variables, however, maintain a similar trend both in the presence and in the absence of missing data, probably because they practically do not present missing data. Ultimately, this discussion highlights that the importance of managing missing data is certainly an open challenge for pyrolysis.

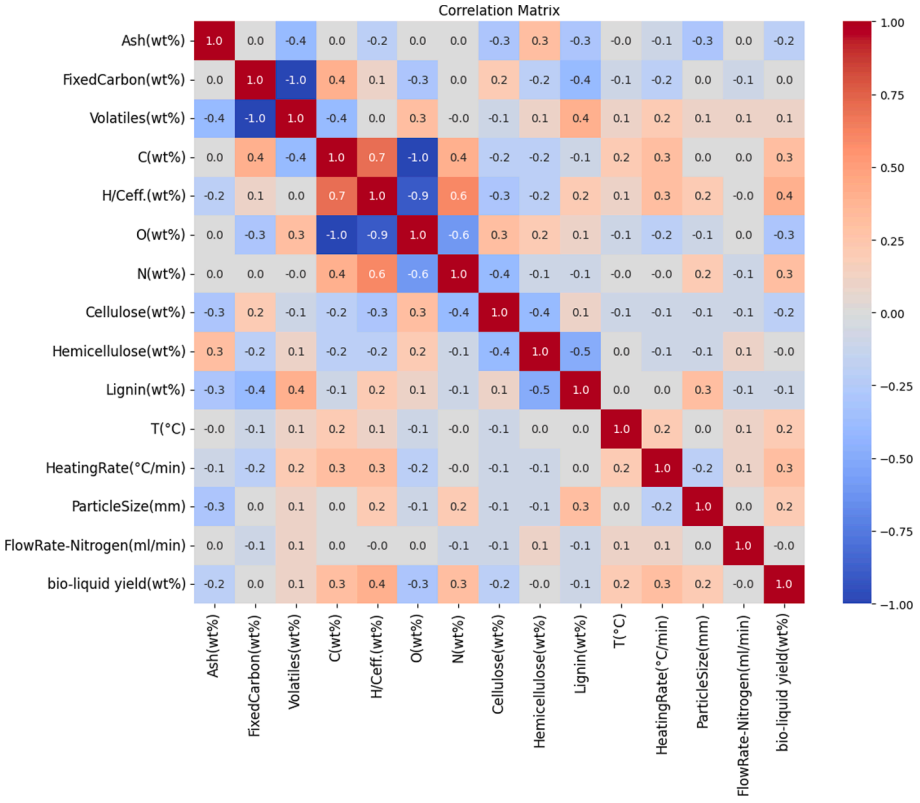


Fig. A.5. Correlation matrix of the reduced dataset without missing values.

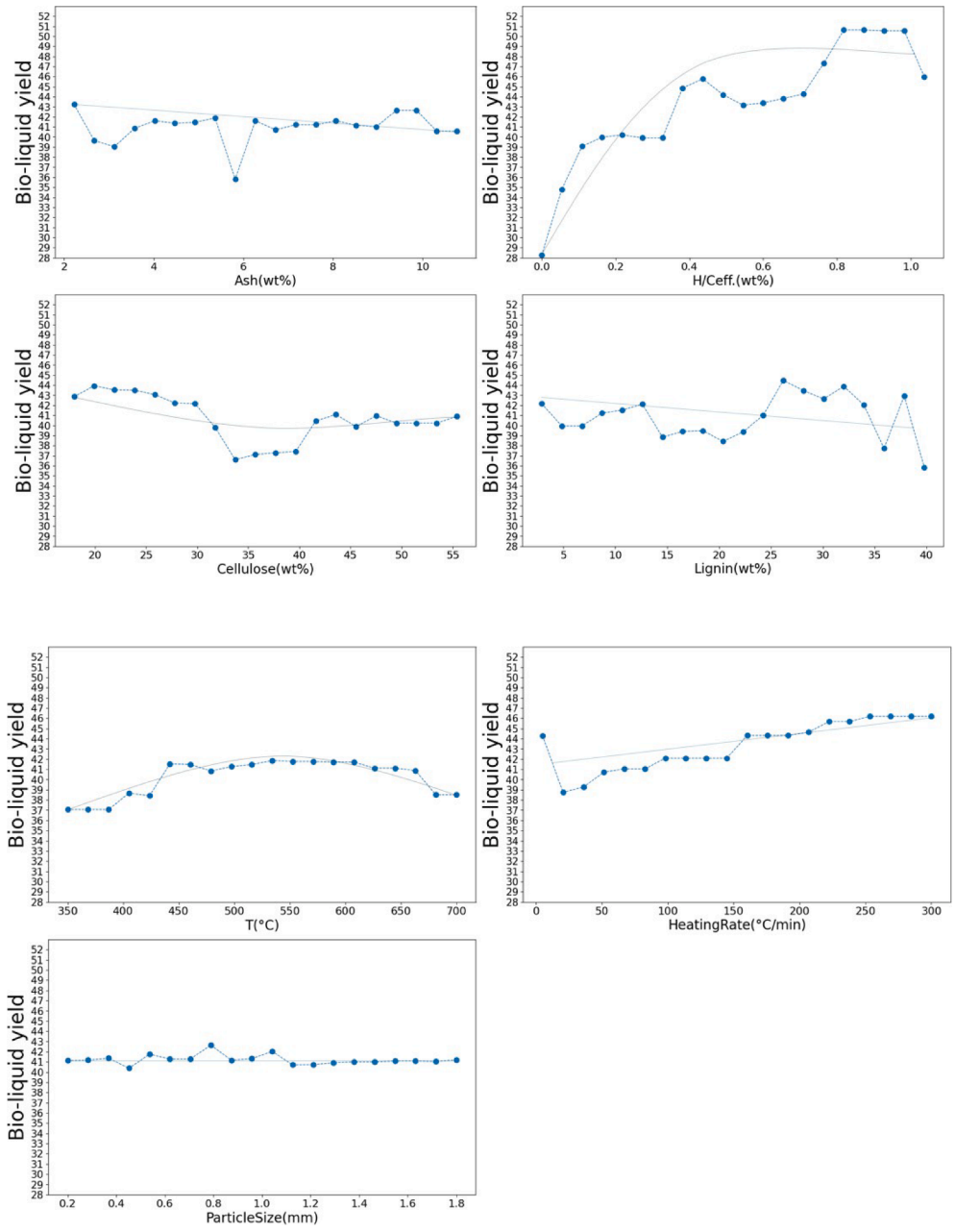


Fig. A.6. Partial dependency plots (PDP) for bio-liquid yield developed on the original dataset with missing data completed with GAIN with respect to biomass properties and operating conditions. The solid line is only a guide-to-eye.

Appendix B

Table B.5 shows a classification of the biomass for the family from a taxonomy point of view, specifying the relative abundance and the different biomass types in the collected dataset. A sample of the dataset and the relative abundance of the different features in the dataset are also reported respectively in Tab. B.6 and Tab. B.7.

Table B5
Classification by family and relative percentage of abundance of the biomasses present in the dataset.

Family (Taxonomy)	%	Biomass
Aceraceae	0.6 %	maple fruit
Anacardiaceae	0.4 %	pistachio shell and seed
Apiaceae	2.5 %	ferula orientalis l.
Arecaceae	1.9 %	oil palm tree waste
Asphodelaceae	2.4 %	eremurus spectabilis
Asteraceae	14.4 %	safflower seed, cirsium arvense, sunflower bagasse, o. acanthium l, onopordum acanthium, xanthium strumarium
Betulaceae	11.5 %	hornbeam shell, hazelnut bagasse, hazelnut cupula, hazelnut shells
Boraginaceae	0.2 %	anchusa azurea stalks
Brassicaceae	6.4 %	rapeseed, rapeseed oil cake
Calophyllaceae	0.2 %	calophyllum inophyllum cake, mesua ferrea seed
Chenopodiaceae	0.2 %	sugar beet bagasse
Chlorellaceae	0.2 %	chlorella vulgaris
Euphorbiaceae	2.8 %	euphorbia macroclada, euphorbia rigida, babool seedsjatropha curcas cake
Eustigmataceae	0.1 %	nannochloropsis
Fabaceae	6.1 %	soybean cake, soybean oil cake, glycyrrhiza glabra l, liquorice, pongamia glabra seed
Lauraceae	2.6 %	laurel, laurel extraction residues, avocado seeds
Linaceae	0.3 %	flax straw, linseed
Malvaceae	0.3 %	jute dust, cotton stalk
Mimosaceae	2.6 %	acacia cincinnata, acacia holosericea, acacia mangium, acacia nilotica
Oleaceae	6.7 %	oilseed, olive bagasse, olive cake, olive residue
Pedaliaceae	2.8 %	sesame stalk
Phormidiaceae	0.3 %	lacustrine alga, spirulina sp
Pinaceae	4.7 %	pine barks, pine chips, pine nedless, white-pine
Poaceae	11.1 %	arundo donax, oat straw, napier grass, lemon grass, miscanthus, rice husk, rice straw, switchgrass, bamboo, moso bamboo, sugarcane bagasse, wheat straw, corn stalks, corncob
Posidoniaceae	0.1 %	posidonia oceanica
Punicaceae	1.1 %	pomegranate seeds
Ranunculaceae	1.3 %	black cumin, black cumin seed cake
Rosaceae	3.3 %	almond shell waste, apricot seed kernel, apricot kernel shell, apricot pulp, cherry seed, peach pulp
Rutaceae	1.6 %	orange bagasse
Sapotaceae	1.5 %	mahua seed, manilkara zapota seed
Scenedesmaceae	1.1 %	s. dimorphus
Scrophulariaceae	1.7 %	paulownia wood
Solanaceae	3.0 %	tobacco residues, potato skin
Theaceae	0.1 %	tea waste
Vitaceae	4.1 %	grape bagasse

Table B6
Sample of the dataset.

Ash(wt %)	FixedCarbon (wt%)	Volatiles (wt%)	C(wt %)	H(wt %)	O(wt %)	N (wt %)	Cellulose (wt%)	Hemicellulose (wt%)	Lignin (wt%)	T (°C)	HeatingRate (°C/min)	ParticleSize (mm)	FlowRate-Nitrogen (ml/min)	yield (wt %)	O-Biooil (wt%)	H-biooil (wt%)	Aqueous phase	Ref.
5.19	5.19	89.63	53.10	6.20	39.90	0.80	42.20	19.40	34.20	550	300.00	0.45	100.00	39.40	19.70	9.10	13.20	[64]
7.84	26.94	65.22	39.34	5.81	53.30	1.54	43.16	30.31	17.02	350	30.00	0.65	0.00	29.38	24.90	5.80	25.00	[65]
7.05	17.67	75.28	45.92	6.21	40.09	6.90	28.58	41.40	4.99	400	5.00	2.00	30.00	53.90	18.37	8.48	39.67	[66]
11.69	11.42	76.89	48.97	6.38	41.63	3.02	22.84	44.01	27.61	500	10.00	0.60	100.00	35.42	17.25	8.10	13.51	[67]
1.04	17.18	81.79	47.33	6.37	45.93	0.37	29.57	17.01	47.97	500	50.00	0.50	150.00	42.30	26.50	8.24	16.00	[68]
1.23	16.61	82.16	52.48	7.58	35.30	4.54	32.06	28.59	29.08	400	5.00	2.00	25.00	43.00	21.07	8.59	31.00	[69]
7.26	10.81	81.93	52.43	6.09	40.86	0.62	28.50	32.50	32.60	500	7.00	0.47	100.00	45.91	27.28	8.02	21.91	[70]
7.33	13.82	78.85	52.90	6.30	40.40	0.40	31.20	45.20	18.10	300	7.00	0.50	200.00	37.04	22.90	7.30	22.84	[71]
5.06	20.23	74.71	51.17	7.95	35.11	5.32	37.14	10.44	26.73	450	35.00	0.85	200.00	48.21	10.54	10.35	8.46	[72]
7.93	11.40	80.67	47.40	5.30	45.50	1.80	42.00	14.70	36.00	550	40.00	0.70	100.00	35.46	20.30	8.90	16.96	[73]
3.90	11.90	84.20	42.00	6.10	47.40	0.40	32.00	19.20	18.80	600	6.00	2.00	100.00	49.74	29.40	9.10	12.74	[74]
12.10	14.56	73.35	51.65	6.20	40.10	2.05	16.11	44.25	9.20	550	7.00	0.60	100.00	42.39	25.32	7.92	15.39	[75]
3.96	7.81	88.24	60.81	10.15	25.95	3.09	26.90	27.70	11.80	600	300.00	0.70	50.00	61.96	14.20	11.95	4.46	[76]
1.13	22.07	76.79	44.73	6.12	48.28	0.87	44.25	24.79	22.61	500	50.00	0.70	100.00	51.00	25.21	8.67	21.50	[77]
1.93	14.88	83.19	49.65	7.54	38.13	4.03	26.98	25.52	39.67	400	5.00	3.20	30.00	42.20	25.40	8.21	33.32	[78]

Table B7
Variable percentages in the dataset.

	Variables	%
Proximate analysis	Ash(wt%)	88
	Fixed Carbon(wt%)	88
	Volatiles(wt%)	88
Ultimate analysis	C(wt%)	100
	H(wt%)	99
	O(wt%)	99
	N(wt%)	96
Macro-Components	Cellulose(wt%)	77
	Hemicellulose(wt%)	57
	Lignin(wt%)	61
Pyrolysis Conditions	T(°C)	100
	HeatingRate(°C/min)	100
	ParticleSize(mm)	100
	FlowRate-Nitrogen(ml/min)	100
Pyrolysis Performances	Bio liquid yield(wt%)	93
	O-Bio oil(wt%)	27
	H-bio oil(wt%)	21
	Aqueous phase (wt%)	36

Data availability

The dataset is available to reviewers upon request and will be made public after acceptance.

References

[1] I. E. Agency, 2019, Energy efficiency.
[2] World Economic Forum. Harnessing artificial intelligence to accelerate the energy transition. White Paper; 2021. p. 25.
[3] Jha SK, Bilalovic J, Jha A, Patel N, Zhang H. Renew Energy and Sustainable Energy Reviews 2017;77:297–317. https://doi.org/10.1016/j. rser.2017.04.018.
[4] Lai JP, Chang YM, Chen CH, Pai PF. A survey of machine learning models in renewable energy predictions. Appl Sci 2020;10. https://doi.org/10.3390/app10175975.
[5] Wang S, Dai G, Yang H, Luo Z. Lignocellulosic biomass pyrolysis mechanism: A state-of-the-art review, Progress in Energy and Combus- tion. Science 2017;62:33–86.
[6] Tripathi M, Sahu JN, Ganesan P. Effect of process parameters on production of biochar from biomass waste through pyrolysis: A review. Renew Sustain Energy Rev 2016;55:467–81.
[7] Bridgwater A V. Pyrolysis of solid biomass: Basics, processes and products. In: Energy from Organic Materials (Biomass): A Volume in the Encyclopedia of Sustainability Science and Technology. Second Edition. Springer; 2018. p. 1221–50.
[8] Kang K, Klinghoffer NB, ElGhamrawy I, Berruti F. Thermochemical conversion of agroforestry biomass and solid waste using decentralized and mobile systems for renewable energy and products. Renewable and Sustainable Energy Reviews 2021;149:111372.
[9] Roy P, Dias G. Prospects for pyrolysis technologies in the bioenergy sector: A review. Renewable and Sustainable Energy Reviews 2017;77:59–69.
[10] Hough BR, Beck DA, Schwartz DT, Pfaendtner J. Application of machine learning to pyrolysis reaction networks: Reducing model solution time to enable process optimization. Comput Chem Eng 2017;104:56–63.
[11] Naqvi SR, Tariq R, Shahbaz M, Naqvi M, Aslam M, Khan Z, et al. Recent developments on sewage sludge pyrolysis and its kinetics: Resources recovery, thermogravimetric platforms, and innovative prospects. Computers & Chemical Engineer- ing 2021;150:107325.
[12] Sakheta A, Raj T, Nayak R, O'Hara I, Ramirez J. Improved predic- tion of biomass gasification models through machine learning, Comput- ers & Chemical. Engineering 2024:108834.
[13] Haq ZU, Ullah H, Khan MNA, Naqvi SR, Ahad A, Amin NAS. Comparative study of machine learning methods integrated with genetic algorithm and particle swarm optimization for bio-char yield prediction. Bioresour Technol 2022;363:128008.
[14] Zhu X, Li Y, Wang X. Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Bioresour Technol 2019;288:121527.
[15] Tang Q, Chen Y, Yang H, Liu M, Xiao H, Wang S, et al. Machine learning prediction of pyrolytic gas yield and compo- sitions with feature reduction methods: Effects of pyrolysis conditions and biomass characteristics. Bioresource technology 2021;339:125581.
[16] Singh Y, Singh D, Singh NK, Sharma A, Abd Rahim E, Ran- ganathan A, Palanichamy P, Palamanit A, Kumar S. Production of bio-oil from lychee-based biomass through pyrolysis and maximization of bio-oil yield with statistical and machine learning techniques. Journal of Cleaner Production 2023;413:137472.
[17] Leng E, He B, Chen J, Liao G, Ma Y, Zhang F, et al. Prediction of three-phase product distribution and bio-oil heating value of biomass fast pyrolysis based on machine learning. Energy 2021;236:121401. https://doi.org/10.1016/j.energy.2021.121401.
[18] Ullah Z, Khan M, Raza Naqvi S, Farooq W, Yang H, Wang S, Vo DVN. A comparative study of machine learning methods for bio-oil yield prediction - a genetic algorithm-based features selection. Bioresource Technology 2021;335:125292. https://doi.org/10.1016/j.biortech. 2021.125292.
[19] Tang Q, Chen Y, Yang H, Liu M, Xiao H, Wu Z, et al. Prediction of bio-oil yield and hydrogen contents based on ma- chine learning method: Effect of biomass compositions and pyrolysis conditions. Energy Fuel 2020;34:11050–60. https://doi.org/10.1021/acs. energyfuels.0c01893.
[20] Yang K, Wu K, Zhang H. Machine learning prediction of the yield and oxygen content of bio-oil via biomass characteristics and pyrolysis conditions. Energy 2022;254:124320. https://doi.org/10.1016/j.energy.2022.124320.
[21] Zhang T, Cao D, Feng X, Zhu J, Lu X, Mu L, et al. Machine learning prediction of bio-oil characteristics quantitatively relating to biomass compositions and pyrolysis conditions. Fuel 2022;312:122812. https://doi.org/10.1016/j.fuel.2021.122812.
[22] Velidandi A, Gandam PK, Chinta ML, Konakanchi S, reddy Bhavanam A, Baadhe RR, Sharma M, Gaffey J, Nguyen QD, Gupta VK. State-of-the-art and future directions of machine learning for biomass characterization and for sustainable biorefinery. Journal of En- ergy Chemistry 2023;81:42–63.
[23] Ortiz M. Biomass pyrolysis dataset. Mendeley Data 2021.
[24] Ullah Z, Naqvi SR, Farooq W, Yang H, Wang S, Vo DVN, et al. A comparative study of machine learning methods for bio-oil yield prediction–a genetic algorithm-based features selection. Biore- source Technology 2021;335:125292.
[25] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM 2020;63:139–44.
[26] Yoon J, Jordon J, Schaar M. Gain: Missing data imputation using generative adversarial nets, in: International conference on machine learning. PMLR 2018:5689–98.
[27] Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in r. J Stat Softw 2011;45:1–67.
[28] Stekhoven DJ, Bühlmann P. Missforest—non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28:112–8.
[29] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for dna microarrays. Bioinformatics 2001;17:520–5.
[30] James G, Witten D, Hastie T, Tibshirani R, et al. An introduction to statistical learning, 112. Springer; 2013.
[31] Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2. Springer; 2009.
[32] Chen T, Guestrin C. Xgboost: A scalable tree boosting system, in. In: Proceedings of the 22nd acm sigkdd international conference on knowl- edge discovery and data mining; 2016. p. 785–94.
[33] Aggarwal CC, et al. Data mining: the textbook, 1. Springer; 2015.
[34] Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, Dean J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, arXiv preprint arXiv:1701.06538 (2017).
[35] Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal 2012;24:69–71.

[36] Su S, Wang J. Machine learning prediction of contents of oxygenated components in bio-oil using extreme gradient boosting method under dif- ferent pyrolysis conditions. Bioresour Technol 2023;379:129040.
[37] Dong Z, Bai X, Xu D, Li W. Machine learning prediction of pyrolytic products of lignocellulosic biomass based on physicochemical characteristics and pyrolysis conditions. Bioresour Technol 2023;367:128182.
[38] Ullah H, Haq ZU, Naqvi SR, Khan MNA, Ahsan M, Wang J. Optimization based comparative study of machine learning methods for the prediction of bio-oil produced from microalgae via pyrolysis. Journal of Analytical and Applied Pyrolysis 2023;170:105879.
[39] Molnar C. Interpretable machine learning. Lulu com 2020.
[40] Zhang H, Cheng YT, Vispute TP, Xiao R, Huber GW. Catalytic conversion of biomass-derived feedstocks into olefins and aromatics with zsm-5: the hydrogen to carbon effective ratio. Energy & Environmental Science 2011;4:2297–307.
[41] Lundberg SM, Lee SI. A unified approach to interpreting model pre- dictions. Advances in neural information processing systems 2017;30.
[42] Ferraro A, Galli A, Moscato V, Sperlì G. Evaluating explainable artificial intelligence tools for hard disk drive predictive maintenance. Artificial Intelligence Review 2023;56:7279–314.
[43] Lundberg SM, Erion GG, Lee SI. Consistent individualized feature attribution for tree ensembles, arXiv preprint arXiv:1802.03888 (2018).
[44] Leng L, Li T, Zhan H, Rizwan M, Zhang W, Peng H, et al. Machine learning-aided prediction of nitrogen heterocycles in bio-oil from the pyrolysis of biomass. Energy 2023;127967.
[45] Bridgwater AV. Review of fast pyrolysis of biomass and product upgrading. Biomass Bioenergy 2012;38:68–94.
[46] Leng E, He B, Chen J, Liao G, Ma Y, Zhang F, et al. Prediction of three-phase product distribution and bio-oil heating value of biomass fast pyrolysis based on machine learning. Energy 2021;236:121401.
[47] Degnan Jr T. Liquid fuel from carbohydrates. Chem Tech 1986:506–11.
[48] Cheng YT, Huber GW. Production of targeted aromatics by using diels–alder classes of reactions with furans and olefins over zsm-5. Green Chemistry 2012;14: 3114–25.
[49] Zhang H, Xiao R, Nie J, Jin B, Shao S, Xiao G. Catalytic pyrolysis of black-liquor lignin by co-feeding with different plastics in a fluidized bed reactor. Bioresour Technol 2015;192:68–74.
[50] Dorado C, Mullen CA, Boateng AA. Origin of carbon in aro- matic and olefin products derived from hzsm-5 catalyzed co-pyrolysis of cellulose and plastics via isotopic labeling. Applied Catalysis B: Envi- ronmental 2015;162:338–45.
[51] Collard FX, Blin J. A review on pyrolysis of biomass constituents: Mechanisms and composition of the products obtained from the conver- sion of cellulose, hemicelluloses and lignin. Renewable and Sustainable Energy Reviews 2014;38: 594–608.
[52] Yogalakshmi K, Sivashanmugam P, Kavitha S, Kannah Y, Varjani S, AdishKumar S, Kumar G, et al. Lignocellulosic biomass-based pyrolysis: A comprehensive review. Chemosphere 2022;286:131824.
[53] Troiano M, Ianzito V, Solimene R, Ganda ET, Salatino P. Fluidized bed pyrolysis of biomass: a model-based assessment of the relevance of heterogeneous secondary reactions and char loading. Energy & Fuels 2022;36:9660–71.
[54] Troiano M, Ianzito V, Solimene R, Ganda ET, Salatino P. Modelling fast pyrolysis of biomass in a fluidized bed reactor. Can J Chem Eng 2023;101:110–20.
[55] Li Y, Gupta R, Zhang Q, You S. Review of biochar production via crop residue pyrolysis: Development and perspectives. Bioresour Technol 2023;369:128423.
[56] Messeri L, Crockett M. Artificial intelligence and illusions of under- standing in scientific research. Nature 2024;627:49–58.
[57] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learningbased science. patterns (ny) 2023;4:100804.
[58] Hogg DW, Villar S. Is machine learning good or bad for the natural sciences? CoRR abs/240518095 2024.
[59] Daoutidis P, Lee JH, Rangarajan S, Chiang L, Gopaluni B, Schweidtmann AM, et al. Machine learning in process systems engineering: Challenges and opportunities. Comput Chem Eng 2023:108523.
[60] Damen D, Doughty H, Farinella GM, Fidler S, Furnari A, Kazakos E, Moltisanti D, Munro J, Perrett T, Price W, et al. The epic- kitchens dataset: Collection, challenges and baselines. IEEE Transac- tions on Pattern Analysis and Machine Intelligence 2020;43:4125–41.
[61] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys 2021;3:422–40.
[62] Ren S, Wu S, Weng Q. Physics-informed machine learning methods for biomass gasification modeling by considering monotonic relationships. Bioresour Technol 2023;369:128472.
[63] Ren S, Wu S, Weng Q, Zhu B, Deng Z. Disentangled representation aided physics-informed neural network for predicting syngas composi- tions of biomass gasification. Energy Fuel 2024;38:2033–45.
[64] Gerçel HF, Gerçel Ö. Bio-oil production from an oilseed by-product: fixed-bed pyrolysis of olive cake. Energy Sources, Part A 2007;29:695–704.
[65] Madhu P, Stephen Livingston T, Manickam IN. Fixed bed pyrolysis of lemongrass (cymbopogon flexuosus): Bio-oil production and charac- terization, Energy Sources. Part A: Recovery, Utilization, and Environ- mental Effects 2017;39: 1359–68.
[66] Ucar S, Ozkan AR. Characterization of products from the pyrolysis of rapeseed oil cake. Bioresource technology 2008;99:8771–6.
[67] Ertaş M, Alma MH. Pyrolysis of laurel (laurus nobilis l.) extraction residues in a fixed-bed reactor: Characterization of bio-oil and bio-char. J Anal Appl Pyrol 2010; 88:22–9.
[68] Demiral I, Kul ŞÇ. Pyrolysis of apricot kernel shell in a fixed-bed reactor: Characterization of bio-oil and char. Journal of analytical and applied pyrolysis 2014;107:17–24.
[69] Duman G, Okutucu C, Ucar S, Stahl R, Yanik J. The slow and fast pyrolysis of cherry seed. Bioresour Technol 2011;102:1869–78.
[70] Ateş F, Pütün AE, Pütün E. Pyrolysis of two different biomass samples in a fixed-bed reactor combined with two different catalysts. Fuel 2006;85:1851–9.
[71] Ateş F, Işıkdağ MA. Evaluation of the role of the pyrolysis tem- perature in straw biomass samples and characterization of the oils by gc/ms. Energy & Fuels 2008; 22:1936–43.
[72] Şen N, Kar Y. Pyrolysis of black cumin seed cake in a fixed-bed reactor. biomass and bioenergy 2011;35:4297–304.
[73] Gerçel HF. Bio-oil production from onopordum acanthium l. by slow pyrolysis. J Anal Appl Pyrol 2011;92:233–8.
[74] Imam T, Capareda S. Characterization of bio-oil, syn-gas and bio- char from switchgrass pyrolysis at various temperatures. J Anal Appl Pyrol 2012;93:170–7.
[75] Pütün AE, Onal E, Uzun BB, Ozbay N. Comparison between the “slow” and “fast” pyrolysis of tobacco residue. Industrial Crops and Products 2007;26:307–14.
[76] Onay O. Fast and catalytic pyrolysis of pistacia khinjuk seed in a well- swept fixed bed reactor. Fuel 2007;86:1452–60.
[77] Yorgun S, Yıldız D. Slow pyrolysis of paulownia wood: Effects of pyrol- ysis parameters on product yields and bio-oil characterization. J Anal Appl Pyrol 2015; 114:68–78.
[78] Uçar S, Karagöz S. The slow pyrolysis of pomegranate seeds: The effect of temperature on the product yields and bio-oil properties. Journal of analytical and applied Pyrolysis 2009;84:151–6.