clean technologies

MDPI

Article

Data-Driven Machine Learning Approach for Predicting the Higher Heating Value of Different Biomass Classes

Inioluwa Christianah Afolabi [1], Emmanuel I. Epelle [2], Burcu Gunes [3], Fatih Güleç [4] and Jude A. Okolie [5,6,*]

[1] Department of Pure and Applied Chemistry, Ladoke Akintola University of Technology, Ogbomoso P.M.B. 4000, Nigeria
[2] School of Computing, Engineering & Physical Sciences, University of the West of Scotland, Paisley PA1 2BE, UK
[3] School of Biotechnology and DCU Water Institute, Dublin City University, Glasnevin, D09 NA55 Dublin, Ireland
[4] Advanced Materials Research Group, Faculty of Engineering, University of Nottingham, Nottingham NG7 2RD, UK
[5] St. Peter's College Muenster, Box 40, Muenster, SK S0K 2Y0, Canada
[6] Gallogly College of Engineering, University of Oklahoma, Norman, OK 73019, USA
* Correspondence: jude.okolie@usask.ca

Abstract: Higher heating values (HHV) is a very useful parameter for assessing the design and large-scale operation of biomass-driven energy systems. HHV is conventionally measured experimentally with an adiabatic oxygen bomb calorimeter. This procedure is often time-consuming and expensive. Furthermore, limited access to the required facilities is the main bottleneck for researchers. Empirical linear and nonlinear models have initially been proposed to address these concerns. However, most of the models showed discrepancies with experimental results. Data-driven machine learning (ML) methods have also been adopted for HHV predictions due to their suitability for nonlinear problems. However, most ML correlations are based on proximate or ultimate analysis. In addition, the models are only applicable to either the originator biomass or one specific type. To address these shortcomings, a total of 227 biomass datasets based on four classes of biomass, including agricultural residue, industrial waste, energy crop, and woody biomass, were employed to develop and verify three different ML models, namely artificial neural network (ANN), decision tree (DT) and random forest (RF). The model incorporates proximate and ultimate analysis data and biomass as input features. RF model is identified as the most reliable because of its lowest mean absolute error (MAE) of 1.01 and mean squared error (MSE) of 1.87. The study findings can be used to predict HHV accurately without performing experiments.

Keywords: machine learning; biomass; higher heating value; biofuel; artificial neural network

Citation: Afolabi, I.C.; Epelle, E.I.; Gunes, B.; Güleç, F.; Okolie, J.A. Data-Driven Machine Learning Approach for Predicting the Higher Heating Value of Different Biomass Classes. *Clean Technol.* **2022**, *4*, 1227–1241. https://doi.org/10.3390/cleantechnol4040075

Academic Editors: Yongliang Xie and Shimao Wang

Received: 8 August 2022
Accepted: 18 November 2022
Published: 22 November 2022

1. Introduction

The chemical and energy industries still rely on the relatively cheap and readily available fossil fuels-based feedstock for energy and power generation. However, the challenges of climate change, decreasing oil reserves, environmental pollution, and elevating fossil fuel prices have promoted an interest in other types of feedstock utilization. Alternative fuel sources such as biomass resources are cheap and abundantly available for the production of heat and electricity through the thermochemical or biological conversion processes [1]. At the same time, the use of biomass resources, specifically waste biomass, helps mitigate the challenges related to their disposal. The International Energy Agency (IEA) noted that biomass accounts for up to 14% of the world's primary energy supply [2]. Moreover, the value is projected to increase further due to the significant interest in renewable energy sources.

Biomass can be grouped into several classes, including agricultural residues, woody biomass, energy crops, and industrial wastes [3]. In addition, they can also undergo thermal

treatment (thermochemical) or reactions with microorganisms or enzymes (biological) to produce green fuels and chemicals. However, it is essential to note that efficient biomass utilization requires detailed knowledge of its physicochemical properties. Higher heating value (HHV) is a very useful parameter for assessing the design and large-scale operation of biomass-fueled energy systems. HHV provides an indication of biomass quality as it defines the optimum energy recovery potential of the biomass. Compared to the lower heating value (LHV), the HHV includes the latent heat of the vaporization of water in the fuel and the products of the reaction [3].

The HHV values of biomass can be experimentally measured with an adiabatic oxygen bomb calorimeter according to the American Society for Testing and Materials (ASTM)-D5865-13 methods [4]. Although experimental HHV determination is a relatively simple and accurate procedure, the experimental facilities might not always be accessible to researchers. In addition, the results could be prone to several experimental errors. Moreover, experiments are also time–consuming and expensive. To address these challenges, several empirical correlations have been proposed for the estimation of HHV values of different biomass materials.

Empirical correlations of HHV values are based on either the proximate or ultimate analysis. Dulong was one of the earliest researchers to present a correlation for the HHV values based on the ultimate analysis data of coal [5]. Other researchers such as Friedl et al. [6], Demirbas [7], and Sheng and Azevedo [8] have also proposed empirical HHV correlations based on the experimental ultimate analysis data. Although the empirical models based on ultimate analysis are promising, most models also show discrepancies with experimental results, as indicated in Figure 1. Therefore, some studies have explored the use of either proximate or ultimate analysis results for HHV predictions.

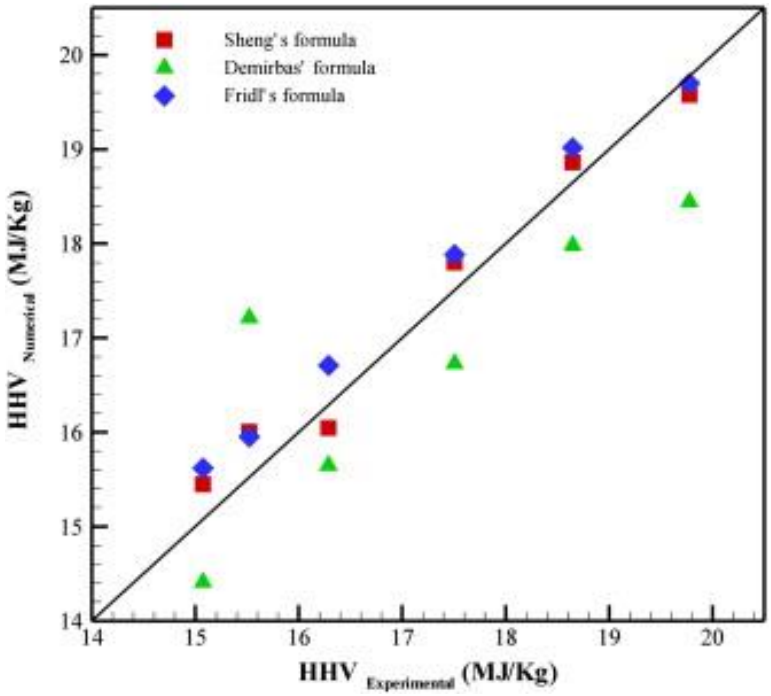

Figure 1. Differences between experimental and model HHV values. Reprinted/adapted with permission from Vaezi et al. [9].

Yu et al. [10] developed an empirical correlation based on proximate analysis using stepwise regression. In their model, the only variables considered were volatile matter (VM) and fixed carbon (FC). In another study, two empirical correlations for HHV value estimation were based on the proximate and ultimate analysis data. However, their model is only applicable to biochars [11]. Choi et al. [12] proposed an empirical model based on

proximate and ultimate analysis. However, the model is only applicable to livestock manure. Kieseler et al. [13] proposed an equation for char production from biomass torrefaction. In summary, most of the empirical models are either applicable to one feedstock or to experimentally measured proximate and ultimate analysis data. Since the HHV values of biomass are highly dependent on biomass characteristics, soil, and climatic conditions, developing correlations that reflect different classifications of biomass is essential.

Data-driven machine learning (ML) has been shown to provide improved accuracy in predicting HHV values of different types of biomass materials [14]. Compared to empirical models, ML methods provide improved accuracy and can also be used to estimate the relative contribution of each input variable on the predicted HHV values. ML methods have been applied for HHV predictions of different types of biomass [13–16]. However, a robust ML model considering different biomass classifications is scarcely reported. In addition, studies comparing the relative prediction capability of decision tree (DT) methods and artificial neural network (ANN) are seldomly reported.

ANN methods are suitable for modeling input-output relationships of complex systems. In addition, they can learn and generalize from the examples provided during the training phase [16]. On the other hand, DT methods are suitable for complex datasets as well as linear and nonlinear problems [17]. Therefore, a comparative evaluation of the performance of DT and ANN methods is invaluable.

Many studies have been reported on the application of ML for HHV predictions. Xing et al. [18] adopted three different ML algorithms (ANN, SVM, and RF) to estimate the HHV of biomass based on their proximate and ultimate analysis. The authors used R^2 as a selection criterion for comparing the accuracy of the models. RF algorithm showed the best performance with $R^2 > 0.94$. However, biomass classification was not included in the ML model. This study presents a comprehensive predictive ML model based on DT, random forest regression (RF), and ANN for predicting HHV values of different biomass classes. Proximate and ultimate analysis data of four different biomass classes, including agricultural residue, industrial waste, energy crop, and woody biomass, were used in the development of the model. Results from the ML models were compared with empirical models and literature-reported ML models.

Relevant Literature and Study Novelty

The application of ML for optimization and prediction has been well documented in several fields ranging from catalytic applications to green hydrogen production, cheminformatics, environmental sustainability, and green chemistry [19,20]. Several researchers have also applied ML models for predicting the HHV of different classes of biomass-related materials and also compounds with varying functional groups. Studies in the literature related to ML applications for HHV predictions can be classified as biomass HHV values predictions based on proximate and ultimate analysis or thermochemical conversion products (e.g., bio-oil, biochar, hydrochar, etc.) HHV predictions.

Leng et al. [21] applied the proximate and ultimate analysis data as well as the biomass composition and particle size for predicting the HHV values of three pyrolysis products (gases, bio-oil, and biochar). The authors also compared the performance of several ML algorithms such as gradient boosting, RF, Support vector machine, and multilayer perceptron algorithms. Among all the evaluated models, RF was reported as the most suitable model. Li et al. [22] showed that ML could be used to optimize the properties of hydrochar (HHV, proximate and ultimate analysis) produced from the hydrothermal carbonization of municipal waste. Recently, the extreme gradient boosting algorithm was adopted for the prediction of biocrude yield and HHV from the hydrothermal liquefaction of wet biomass [23]. The ML-based algorithm was developed based on the physicochemical properties of the feedstock as well as hydrothermal liquefaction operating conditions. The model showed improved performance in HHV predictions with an R^2 value > 0.9.

García Nieto et al. [15] developed a ML-based predictive model to estimate the HHV during the torrefaction process. Their results showed the importance of each physical-

chemical variable on the HHV values of different biomass materials. However, the proposed model only applies to torrefied biomass materials with little information on the biomass precursor. Ighalo et al. [14] employed the linear regression algorithm and stochastic gradient descent algorithm to predict the HHV value of biomass. A mean absolute error (MAE) of 6.823 was obtained for the linear regression algorithm, while the stochastic gradient algorithm had a superior MAE of 13.87. Although the authors adopted the combination of proximate and ultimate analysis data for the ML model, the lack of sufficient data limits the accuracy of the ML model. Recently Taki et al. [16] developed an ML model for predicting HHV values of municipal solid waste without considering other biomass materials. Considering that previous studies related to the ML-based prediction of HHV are focused on one class of biomass or exclude biomass class as part of the feature variables, it is imperative to include different biomass classes as part of the ML models. Recently, Gulec et al. [24] adopted the ANN for the prediction of HHV for several biomass feedstocks. The authors evaluated the impact of activation functions, algorithms, hidden layers, dataset, and randomization of the dataset on HHV prediction accuracy.

Despite utilizing ML-based models for HHV predictions mentioned above, different feedstock classifications, integrating models that comprise proximate and ultimate analysis, and biomass classification have not been comprehensively considered. Furthermore, there is a lack of discussion on models constructed by different ML algorithms and several input feature combinations. The present study attempts to fill the knowledge gaps. The difference between the methodology presented in this paper and previous studies are the following:

- This study proposes a comprehensive ML model comprising proximate and ultimate analysis and different biomass classification input features. Specifically, the biomass classification is selected to capture a wide range of materials, including agricultural residues, energy crops, woody biomass, and industrial waste.
- This study applies a robust data set of 227 different biomass materials and computationally compares the performance of three different ML algorithms, including RF, DT, and ANN.

2. Methodology

2.1. Dataset Collection and Pre-Processing

In this work, three machine learning algorithms were compared to predict the higher heating value of four biomass classes: agricultural residues, energy crops, woody biomass, and industrial waste. The ML algorithms include tree-based ML models, i.e., decision trees, random forests, and neural networks (artificial neural networks). The datasets containing 227 instances of different biomass types studied were sourced from publications and compiled in an excel spreadsheet (see Supplementary Materials). Although some of the curated data had missing values, these were replaced with the average of the observation for the sake of model training. Furthermore, the dataset was split into features (fixed carbon (FC), volatile matter (VM), ash (AC), carbon (C), hydrogen (H), nitrogen (N), sulphur (S), and oxygen(O)) and target higher heating value (HHV).

The features were normalized between 0 and 1 to avoid excessive dominance of larger values on smaller ones and to enhance fast convergence during model training using Equation (1), afterwards transformed back to the original values.

$$X_{norm} = \frac{x_o[n] - x_o[min]}{max(x_o) - x_o[min]} \quad (1)$$

where X_{norm} is the normalized x feature, $x_o[n]$ is the nth sample of the original dataset, $x_o[min]$ and $max(x_o)$ are the minimum and maximum values of original features. Schematics of the data pre-processing stage are outlined in Figure 2. The four biomass classes studied as category features were one-hot encoded to 0 s and 1 s with python software to comprehend the ML models easily. The processed datasets were divided randomly into

70% training and 30% testing sets. The choice of this division ratio agrees with previous literature on machine learning [15,17,18].

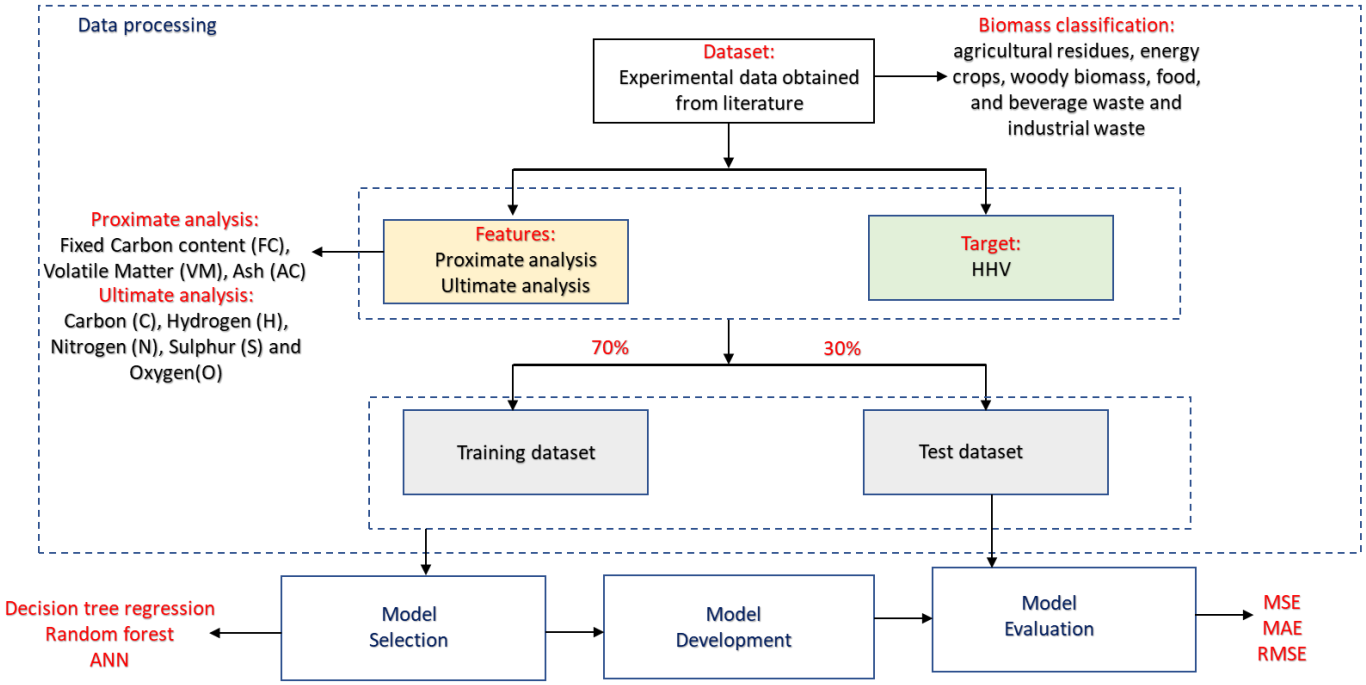

Figure 2. An overview of the methodology used in the study.

Since this study focused on supervised learning, the models were trained with the training datasets' features and their corresponding targets. To check the model's generalization ability, 30% of testing datasets that were not used during model training were used to evaluate the model's performance. The models' performance was assessed with mean squared error (MSE), root mean square error (RMSE), and mean absolute error (MAE). The MSE and MAE values are calculated from Equations (2) and (3), respectively, where 'n' represents the total number of samples [24]. The MSE is the mean square difference between the actual and predicted values [17].

On the contrary, the MAE indicates the absolute sum of the differences between the actual and predicted output divided by the total number of outcomes. The RMSE is the square root of the MSE. Since some of the datasets contain missing values or values not reported in the extrapolated publication, the MSE is a suitable criterion for assessing the accuracy of the ML model. Lower values of both MAE and MSE indicate improved ML model performance and predictability. MAE is very sensitive to outliers and often performs poorly when the data contains significant errors. It should be noted that the coefficient of determination (R^2) was not considered as a ML model evaluation criterion in this study due to the nonlinear relationships between the target and features.

The following section presents a brief overview of the ANN and decision tree models employed in this study.

$$\text{MSE} = \frac{1}{n} \sum_{n-1}^{n} (\text{Experimental values} - \text{Predicted values}) \quad (2)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\text{Experimental value} - \text{predicted value}| \quad (3)$$

2.2. Overview of the Machine Learning Algorithm

2.2.1. Artificial Neural Networks (ANN)

ANN represents one of the most advanced ML algorithms suitable for evaluating the relationship between features and targets. Its self-adaptive nature and excellent generalization ability in predicting the outputs of new datasets are the main advantages [25]. ANN has a universal function approximator that improves the approximation of complex nonlinear problems [19,20].

The ANN architecture used in this study is a four-layered model with one input layer, two hidden layers and one output layer, as depicted in Figure 3. The input layer consists of 9 hidden neurons from the input features (fixed carbon, volatile matter, ash, carbon, hydrogen, nitrogen, sulphur, and oxygen). The two hidden layers consist of 128 and 32 hidden neurons, respectively. It is worth noting that the hidden neurons are selected based on trial-and-error methods, while the output layer has one neuron from the target variable (higher heating value). The input features were fed into the model, processed with the help of processing units, otherwise known as the neurons in the hidden layers, then transferred to the output layer with the help of the activation function in the forward propagation. Mean square error was used as a loss function to determine the distance of the predicted value from the actual target. This distance was then adjusted with the help of the 'adam' optimizer, which adjusts the weight and bias supplied from the input-hidden layer until the gap between the predicted and actual target is reduced to the barest minimum in the backpropagation. The training stops whenever the network parameters are adjusted to the minimum value. The epoch number was varied between 10 and 50, and the model was closely monitored for overfitting. Ten epochs were selected for this study because the model started overfitting beyond ten.

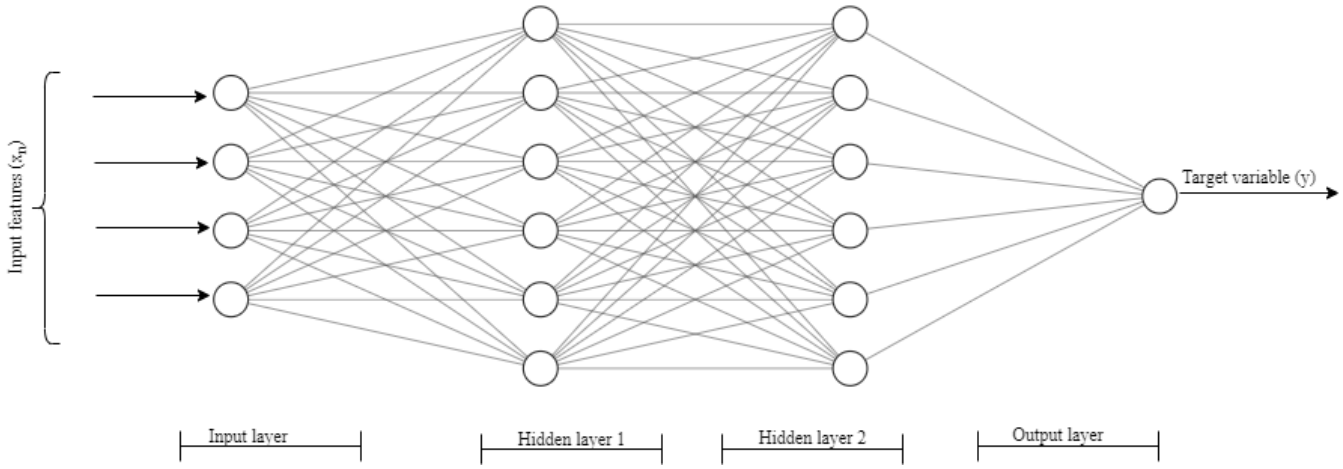

Figure 3. The schematic representation of a four-layered ANN model.

The trained model was then tested on a new set of unknown data to check if the model had good generalization ability or if data had been memorized during the training process. The result of the model on the testing dataset is presented in the Section 3 showing the excellent generalization ability of the developed ANN model.

2.2.2. Decision Tree Regression (DT)

A decision tree regression (DT) is a tree-based supervised machine learning algorithm for predicting target variables. This non-parametric algorithm divides a population of training datasets into branch-like segments in an inverted tree form with root, internal, and leaf nodes. The leaves are the decisions or the outcomes. The decision nodes are where the data is split. The algorithm uses a binary tree to recursively divide the output space into sub-sets where the output distribution is more homogeneous in succession, then tries to reduce the prediction error in separate parts [17]. This process is replicated for each new

branch. It is worth noting that the prediction of continuous output variables, such as HHV, in these studies is achieved by some sets of logical rules [26].

The DT is often used because of its low computational power and strength in dealing with large datasets without imposing complicated parametric structures. When a large dataset is present, training datasets could be divided into training and validation sets. The training sets are then used to build the tree model, while the validation set decides the appropriate tree size needed to achieve the optimal final model. The working principle of this algorithm is fully discussed in the excellent review by Umenweke et al. [17]. Some subsequent studies have also explored DT in predicting the higher heating values of diverse biomass [26,27]. Compared to other ML models, DT can produce an ML model that has the potential to be represented as logical statements and a set of rules [27].

2.2.3. Random Forest (RF)

A random forest (RF) is a form of tree-based ensemble ML algorithm used for predicting continuous values. It operates by constructing multiple decision trees for training features and obtaining the target values as an average prediction of each tree [28]. The training datasets, which consist of 'n' total samples and 'm' features, are subjected to bootstrap sampling. In the sampling, a number of sample sets (n) are randomly generated from the original training datasets. Then, for each sample set, the samples are randomly divided into one-third in-of-bag samples and two-thirds out-of-bag samples. The in-of-bag samples are used for the training, while the out-of-bag samples are used to determine the optimal number of trees for the task. The prediction of the RF-trained model is the average prediction of all the trees in the model. Further details of the theory of random forest are discussed elsewhere [17].

2.3. Empirical Correlations

There are several empirical correlations for predicting the HHV of biomass feedstock based on the proximate analysis and ultimate analysis. Also, the correlations could be linear or nonlinear. Table 1 provides different empirical models whose results are compared with the ML model in this study.

Table 1. Overview of different empirical correlations compared with ML predictions.

Empirical Correlation	Equation for HHV in MJ/kg	Biomass Used	References
Demirbas correlation	0.01 (33.5C + 142.3H − 15.4O − 14.5 N)	Agricultural residues	Demirbaş [7]
Sheng correlation	−1.3675 + 0.3137C + 0.7009H + 0.0318O	Agricultural residues	Sheng and Azevedo [8]
Friedl correlation	20600 + 3.55 C^2 − 232C − 2230H + (51.2C × H) +131N	Wood, grass, rye, rape, reed, brewery waste, and poultry litter	Friedl et al. [6]
Yu correlation (Ultimate analysis)	0.2949C + 0.8250H	Agricultural residues	Yu et al. [10]
Yu correlation (Proximate analysis)	0.1905VM + 0.2521FC	Agricultural residues	Yu et al. [10]
Qian correlation (Ultimate analysis)	32.9C + 162.7H − 16.2O − 954.4S + 1.408	Biochars	Qian et al. [11]
Qian correlation (proximate analysis)	−30.3FC^2 + 65.2Ash^2 + 55.4FC − 48.5Ash + 9.591	Biochars	Qian et al. [11]

3. Results and Discussion

3.1. Statistical Analysis of the Dataset

Statistical analysis of the collected dataset was presented to understand the details and trends of the features and targets as well as their relationship with HHV. The box plot analysis results for the features and targets are outlined in Figure 4. As shown in the box plot results, the values of VM and FC contents vary significantly, with values ranging

from 59.3–95.5 wt.% for VM and 0.1–37.9 wt.% for FC. In addition, the ash (0–27 wt.%), C (34.6–57.7 wt.%), and O (32.1–73.8 wt.%) contents also vary at a relatively high percentage. It should be mentioned that there are some outliers in the data, and they are identified with dots outside the box. The large range of ash, C, VM, O and FC could be attributed to different biomass materials considered in the study. Biomass classes including agricultural residues, energy crops, woody biomass and industrial waste have distinct physico-chemical properties based on their origin and sources of generation.

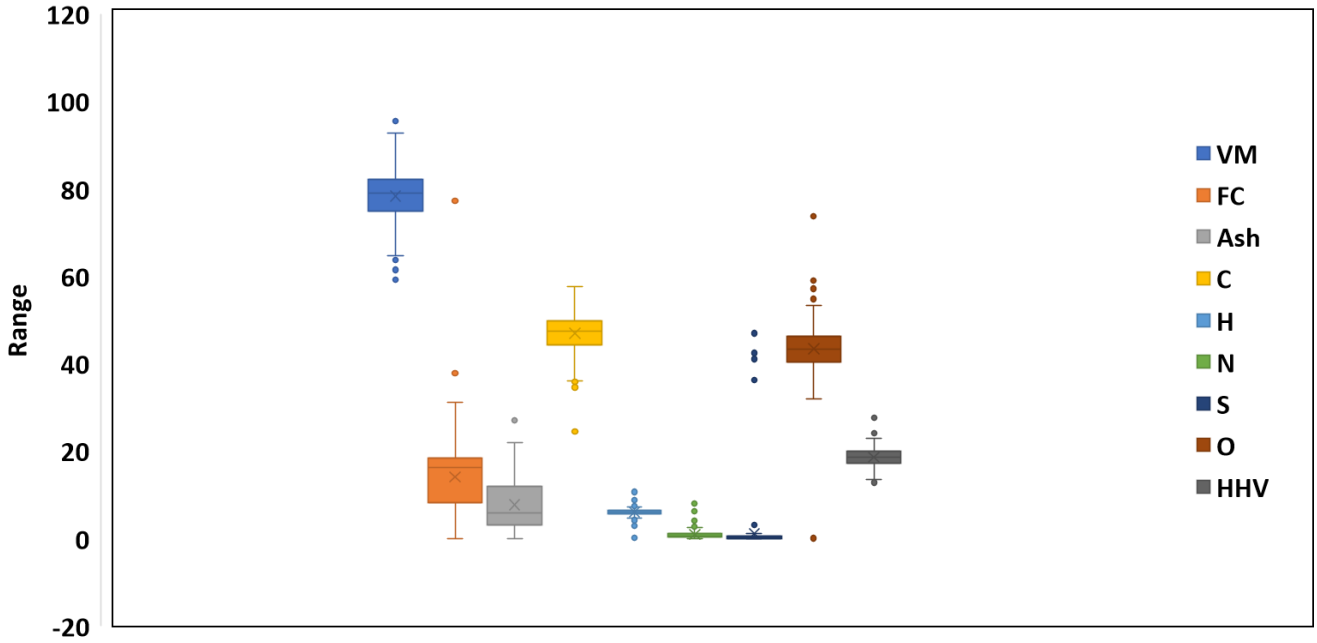

Figure 4. Box plot results of the statistical analysis of the input features and target (HHV) obtained from the entire dataset used in the machine learning models.

The S, N, and H contents fall within narrow ranges of 1.6–3.2 wt.%, 1–8 wt.%, and 0.3–10.7 wt.%, respectively. Furthermore, the output target (HHV) also falls within a narrow range of 12.8–27.8 wt.% including the outlier. It should be mentioned that the diverse range of values for the input features is very useful in data generalization.

It is important to note that when there are no phenomenological (first principles) models, the HHV of biomass materials is predicted using empirical models. Therefore, it is imperative to evaluate the relationship between the features (proximate and ultimate analysis) as well as the target (HHV). Determination of the appropriate relationship is also useful in the selection of a suitable ML model since some regression models are suitable for linear problems. For instance, the DT models are suitable for linear and nonlinear problems while the traditional regression models perform better for linear problems.

Cross-plots were presented from the experimental datasets. The cross plots shown in Figures 5 and 6 compare individual constituents of the proximate and ultimate analyses against their corresponding HHV. The figures show that a linear dependence exists between the HHV and the percentage of C and VM. Moreover, a significant scatter exists in most of the cross-plots including those HHV versus N and FC content. Therefore, it can be inferred that the implementation of linear models may not be the most effective for accurate predictions of HHV values. The use of ML-based models such as DT, RF, and ANN are all suitable for nonlinear relationships. Accordingly, in this study, we have implemented the three models with the same data used for the cross-plots.

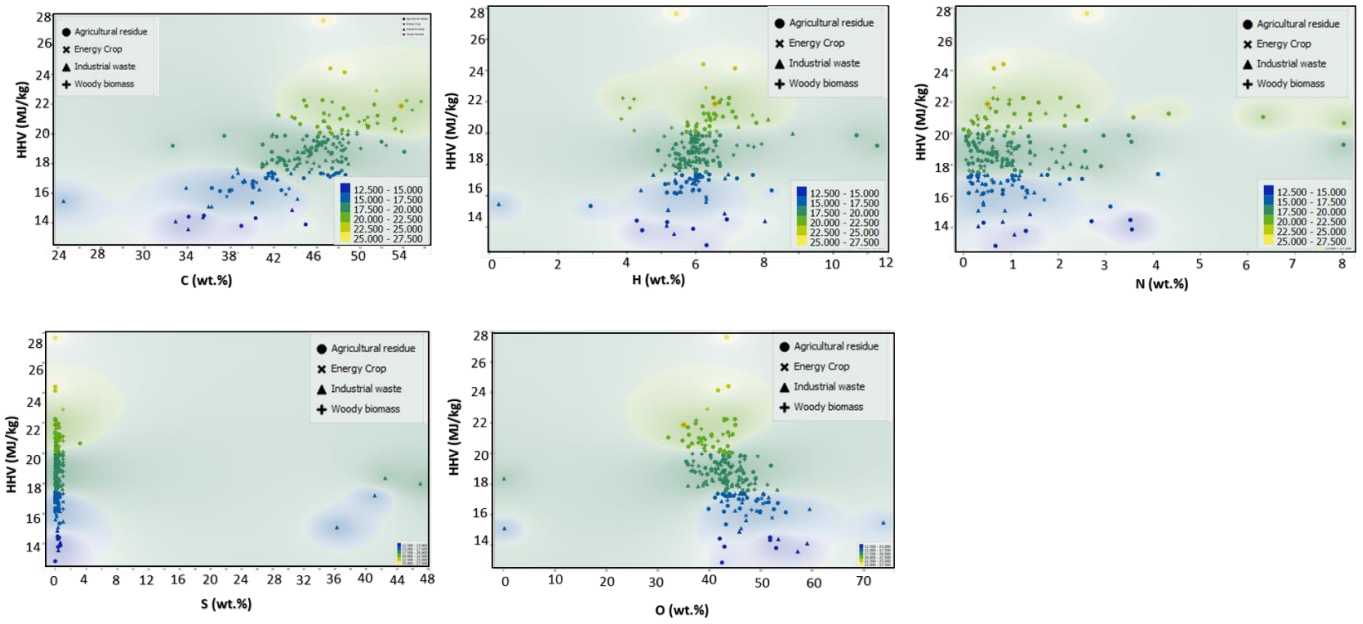

Figure 5. Cross-plots of HHV against the ultimate analysis of different biomass classes.

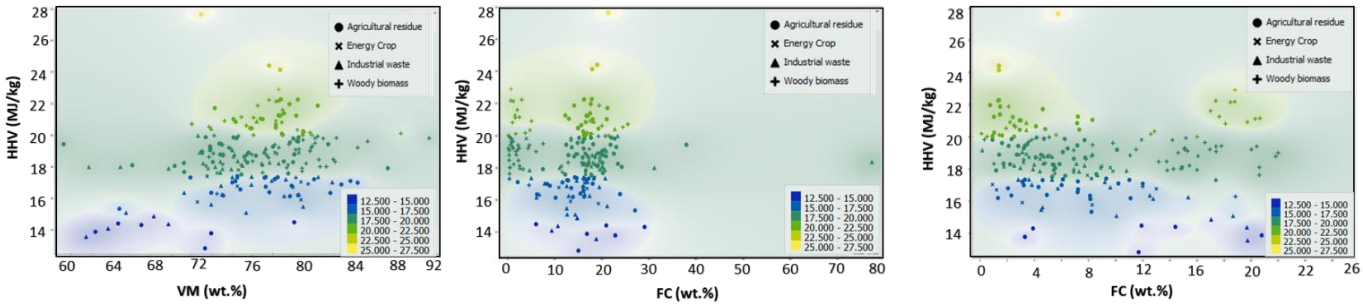

Figure 6. Cross-plots of HHV against the proximate analysis of different biomass classes.

3.2. Model Performance Evaluation

Different statistical quality measurements including MSE, MAE, and RMSE were used to assess the accuracy and preciseness of different ML models used in the study. Model performance evaluations were also assessed by providing a graphical comparison between the predicted HHV from different models. Figure 7 shows the results of graphical difference plots between different models used in the study. Also, the statistical quality measurement criteria are outlined in Table 2. It should be mentioned that the test dataset is used in evaluating the predicted HHV.

As shown in Figure 7, it can be demonstrated that the RF model based on the proximate and ultimate analysis input feature was the most reliable model because the range of differences between the experimentally measured HHV and the predicted HHV is the lowest. The graphical results were also corroborated with data reported in Table 2. The RF model has the lowest MAE (1.01), MSE (1.87), and RMSE (1.37) among the compared ML models. On the contrary, the DT model showed the highest values of MAE = 1.48, MSE = 4.36, and RMSE = 2.09. Based on the results of Figure 5 and Table 1, it can be demonstrated that the accuracy of the ML models is in the order of RF > ANN > DT. The authors referred to Table S1 of the Supplementary Materials for memory utilization comparison.

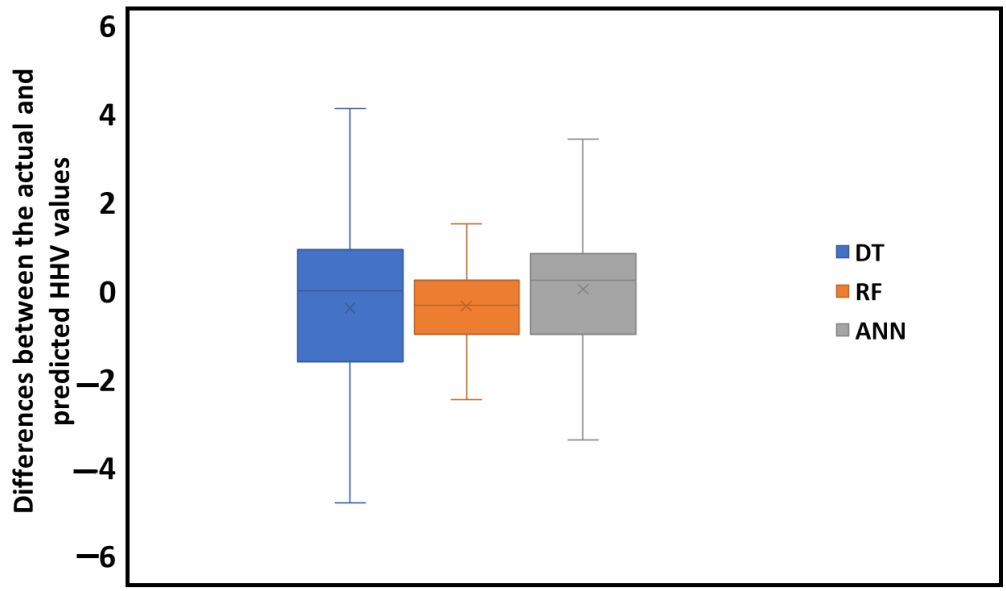

Figure 7. Graphical comparison of experimental and predicted HHV of different ML models using the test dataset.

Table 2. Comparison of the statistical measurement criteria for different machine learning models.

ML Models	MAE	MSE	RMSE
DT	1.48	4.36	2.09
RF	1.01	1.87	1.37
ANN	1.21	2.43	1.56

3.3. Feature Analysis of the Best Model

Several studies have applied ML models to predict the HHV of different types of biomass materials. However, most of the studies apply either proximate or ultimate analysis as input features. Matin and Chelgani [29] compared the best and most reliable input features for ML models. They noted that the ultimate analysis is the most effective input for the RF-ML models. The authors also reported that the use of proximate analysis as an input feature produces satisfactory results. Some other studies have demonstrated that the proximate analysis is a promising input feature for HHV predictions with ML models [12,14,16,30]. However, few studies have accessed the relative importance and contribution of each input feature (including the proximate and ultimate analysis data) towards the ML model accuracy. It is important to assess the relative contribution of each input feature to determine the most important.

A feature evaluation was performed to assess the input feature importance and contribution to the ML model. The RF model was used to perform the feature evaluation since it is the most promising ML model. The feature evaluation approach provides a quantitative evaluation by changing the values of the predictor variables, one at a time, and assessing the decrease in model accuracy for each variable [31]. The exhaustive feature selection method that tries every possible combination of variables and returns the best-performing subset was used. Figure 8 depicts the ranking result based on their importance. It can be demonstrated that the top five input features that contribute significantly to HHV predictions include ash, C, VM, N contents, and the biomass classes. The ash content was the first ranked with 15.6%, while C content is the second-ranked with 12.9%. In contrast, VM, N, H, and biomass classes have a total contribution of 35.3%. The class of biomass has a contribution of 7.8%.

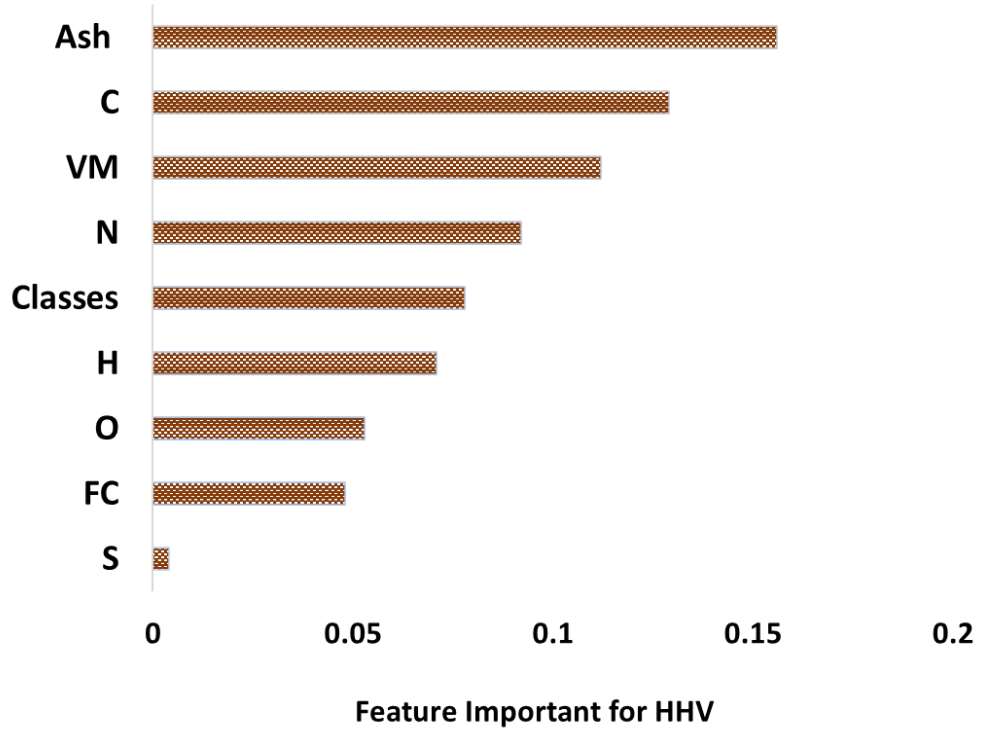

Figure 8. Feature importance of HHV.

The contribution of biomass classes to the HHV prediction model highlights the relevance of this study. While other studies focus on developing ML models for HHV predictions without considering the classes of biomass as input, this study incorporated several classes of biomass as input features. More interestingly, the S content had the least contribution to HHV (<1%).

3.4. Comparison of Other Models from Literature and Empirical Models

The accuracy and validity of ML models were further compared with published literature values as shown in Table 3. However, it should be noted that, while the proposed ML model in this study is dependent on four classes of biomass considered (agricultural residues, energy crops, industrial waste, and woody biomass), the ones in the literature are for the fuels from that they are derived. For those related to solid biomass materials, Ghugare et al. [32] predicted a low MAE of 0.73 with ANN. Using proximate analysis of Chinese and U.S. coal samples as input feedstock, Tan et al. [33] obtained an MAE of 2.16 for Chinese coal and 2.42 for U.S. coal. The authors also proposed the possibility of developing a universal correlation with coal samples from both China and the U.S. by merging data from both coal samples. The MAE values for the universal correlation increased to 6.24. Their results infer that the correlation between the proximate analysis and HHV of coal from different regions is not the same. As a result, we proposed the introduction of biomass classes as part of the input features. Lower values of MAE and MSE from this study show that the introduction of biomass classes as part of the input feature could improve ML model accuracy. It should be noted that some of the reported ML models reported in Table 3 showed higher RMSE values compared to the model reported in this study, therefore it can be inferred that the prediction ability of the model is comparable to those reported in the literature.

Table 3. Comparison of statistical measurement criteria with ML-based HHV values reported in the literature.

Machine Learning Model	Input Feature	RMSE	References
RF	Proximate and ultimate analysis, biomass classes	1.37	This study
DT	Proximate and ultimate analysis, biomass classes	2.09	This study
ANN	Proximate and ultimate analysis, biomass classes	1.56	This study
ANN	Proximate analysis of biochars	0.65	Çakman et al., 2021 [34]
Extreme learning machine	Ultimate analysis	1.93	Dai et al. [35]
ANN	Ultimate analysis	3.87	Xing et al. [18]
RF	Ultimate analysis	2.39	Xing et al. [18]
SVM	Ultimate analysis	2.53	Xing et al. [18]
Genetic programming (GP)	Ultimate analysis	0.95	Ghugare et al. [32]
Multilayer perceptron neural network (MLP)	Ultimate analysis	0.99	Ghugare et al. [32]

The HHV prediction from the proposed RF-ML model was also compared with empirical models for different classes of biomass and the results are presented in Figure 9. Compared to empirical models, the proposed ML model was able to accurately predict the HHV values of industrial wastes and woody biomass with a difference between experimental and model values less than 2%. However, larger differences exist between the ML model and experimental values for agricultural residues and energy crops, although the deviation is less than 5%. Overall, comparing the estimation performance between the RF-ML model and empirical models, it can be observed that the ML model can provide more precise HHV estimation results. The study findings can be used for accurate prediction of HHV without performing experiments.

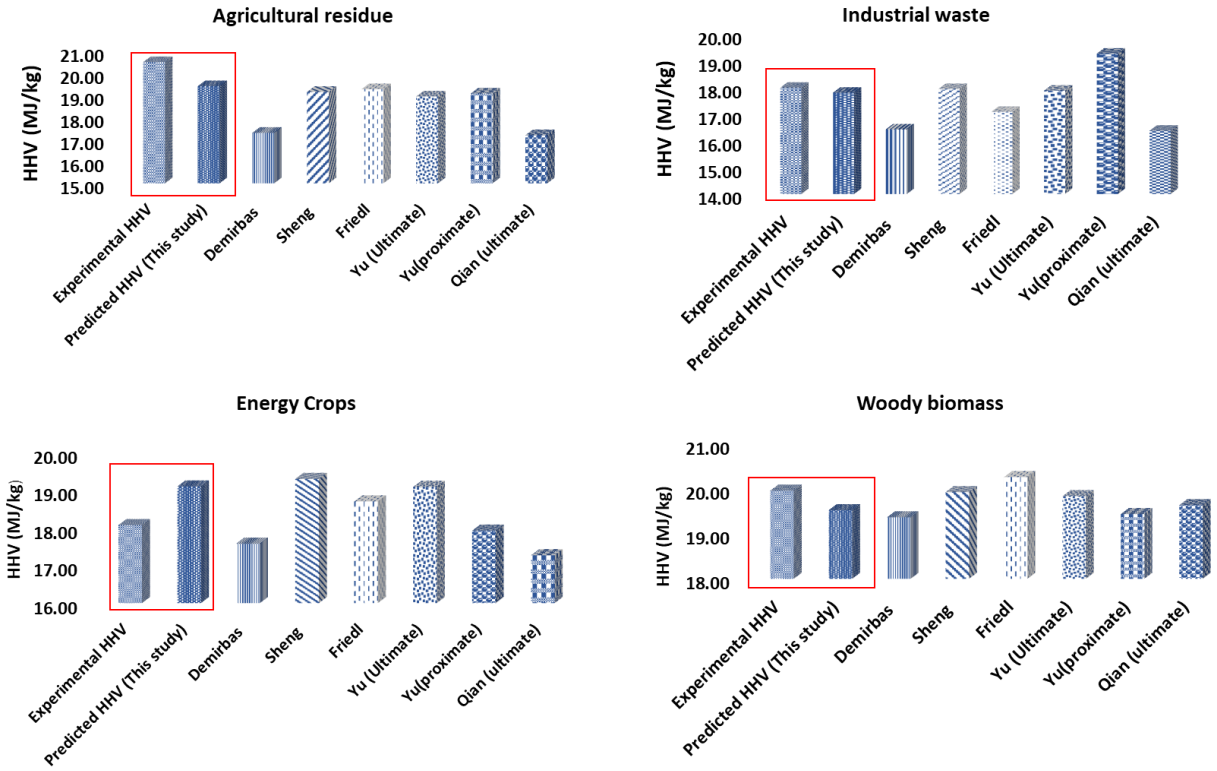

Figure 9. Comparison of the predicted HHV from ML model with empirical models. Note: All the empirical correlations are outlined in Table 1.

4. Conclusions

Although most studies measure the HHV of biomass materials using experimental methods such as bomb calorimeter, several challenges are inherent. For example, experimental methods are often expensive and not always available. In this study, a comprehensive data-driven ML model was developed based on proximate and ultimate analysis data of biomass materials. Another novelty is the consideration of four different types of biomass classes as part of the input features. Model predictions from three different ML models (decision tree (DT), random forest regression (RF), and artificial neural network (ANN)) suitable for nonlinear problems are compared with several statistical measurement criteria. Based on the lowest value of mean squared error of 1.87, mean absolute error of 1.01 and root mean squared error of 1.37 the RF model was selected as the most accurate model. The feature analysis for relative importance shows that both input features from the proximate and ultimate analysis are important. The top five input features that contribute significantly to HHV predictions include ash, C, VM, N contents, and the biomass classes. It is worth noting that the ML model presented in this study also predicts better than the empirical correlations for the biomass classes under consideration. Future studies would focus on developing a comprehensive universal ML model to predict HHV values of coal, torrefied biomass, hydrochar, and biochar.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cleantechnol4040075/s1, Table S1: Memory utilization comparison.

Author Contributions: Conceptualization, F.G. and J.A.O.; Data curation, I.C.A. and F.G.; Formal analysis, I.C.A. and J.A.O.; Funding acquisition, B.G. and J.A.O.; Methodology, I.C.A., E.I.E., B.G. and J.A.O.; Project administration, B.G. and J.A.O.; Resources, E.I.E.; Software, E.I.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Please refer to the following link for the ML algorithm and data. https://colab.research.google.com/drive/1fmgcohRQ5Rnawg7HEkWHrxA9n9c8ygB0?usp=sharing (accessed on 2 August 2022), https://drive.google.com/file/d/19rspU-20v4_qgD-ZTRtpsEyoYtfTQ9G5/view?usp=sharing (accessed on 2 August 2022).

Acknowledgments: The authors would like to thank all the reviewers and editors that have contributed to the success of this work.

Conflicts of Interest: The authors declare that there is no conflict of interest.

Abbreviations

AC	Ash content
ANN	Artificial neural network
ASTM	American Society for Testing and Materials
C	Carbon content
DT	Decision tree
FC	Fixed carbon content
H	Hydrogen content
HHV	Higher heating value
IEA	International Energy Agency
LHV	Lower heating value
MAE	Mean absolute error
ML	Machine learning
MSE	Mean Squared Error
N	Nitrogen content
O	Oxygen content

RF	Random Forest regression
RMSE	Root mean square error
S	Sulphur content
VM	Volatile matter

References

1. Okolie, J.A.; Nanda, S.; Dalai, A.K.; Kozinski, J.A. Hydrothermal gasification of soybean straw and flax straw for hydrogen-rich syngas production: Experimental and thermodynamic modeling. *Energy Convers. Manag.* **2020**, *208*, 112545. [CrossRef]
2. IEA. International Energy Agency: Key World Energy Statistics. *Statistics* **2016**, *78*. Available online: http://www.oecd-ilibrary.org/energy/key-world-energy-statistics-2010_9789264095243-en (accessed on 3 June 2022).
3. Cai, J.; He, Y.; Yu, X.; Banks, S.W.; Yang, Y.; Zhang, X.; Yu, Y.; Liu, R.; Bridgwater, A.V. Review of physicochemical properties and analytical characterization of lignocellulosic biomass. *Renew. Sustain. Energy Rev.* **2017**, *76*, 309–322. [CrossRef]
4. Ohliger, A.; Förster, M.; Kneer, R. Torrefaction of beechwood: A parametric study including heat of reaction and grindability. *Fuel* **2013**, *104*, 607–613. [CrossRef]
5. Channiwala, S.A.; Parikh, P.P. A unified correlation for estimating HHV of solid, liquid and gaseous fuels. *Fuel* **2002**, *81*, 1051–1063. [CrossRef]
6. Friedl, A.; Padouvas, E.; Rotter, H.; Varmuza, K. Prediction of heating values of biomass fuel from elemental composition. *Anal. Chim. Acta* **2005**, *544*, 191–198. [CrossRef]
7. Demirbaş, A. Calculation of higher heating values of biomass fuels. *Fuel* **1997**, *76*, 431–434. [CrossRef]
8. Sheng, C.; Azevedo, J.L.T. Estimating the higher heating value of biomass fuels from basic analysis data. *Biomass Bioenergy* **2005**, *28*, 499–507. [CrossRef]
9. Vaezi, M.; Passandideh-Fard, M.; Moghiman, M.; Charmchi, M. On a methodology for selecting biomass materials for gasification purposes. *Fuel Process. Technol.* **2012**, *98*, 74–81. [CrossRef]
10. Yu, Z.T.; Xu, X.; Hu, Y.C.; Fan, L.W.; Cen, K.F. Prediction of higher heating values of biomass from proximate and ultimate analyses. *Fuel* **2011**, *90*, 1128–1132. [CrossRef]
11. Qian, C.; Li, Q.; Zhang, Z.; Wang, X.; Hu, J.; Cao, W. Prediction of higher heating values of biochar from proximate and ultimate analysis. *Fuel* **2020**, *265*, 116925. [CrossRef]
12. Choi, H.L.; Sudiarto, S.I.A.; Renggaman, A. Prediction of livestock manure and mixture higher heating value based on fundamental analysis. *Fuel* **2014**, *116*, 772–780. [CrossRef]
13. Kieseler, S.; Neubauer, Y.; Zobel, N. Ultimate and proximate correlations for estimating the higher heating value of hydrothermal solids. *Energy Fuels* **2013**, *27*, 908–918. [CrossRef]
14. Ighalo, J.O.; Adeniyi, A.G.; Marques, G. Application of linear regression algorithm and stochastic gradient descent in a machine-learning environment for predicting biomass higher heating value. *Biofuels Bioprod. Biorefin.* **2020**, *14*, 1286–1295. [CrossRef]
15. Nieto, P.J.G.; García-Gonzalo, E.; Lasheras, F.S.; Paredes-Sánchez, J.P.; Fernández, P.R. Forecast of the higher heating value in biomass torrefaction by means of machine learning techniques. *J. Comput. Appl. Math.* **2019**, *357*, 284–301. [CrossRef]
16. Taki, M.; Rohani, A. Machine learning models for prediction the Higher Heating Value (HHV) of Municipal Solid Waste (MSW) for waste-to-energy evaluation. *Case Stud. Therm. Eng.* **2022**, *31*, 101823. [CrossRef]
17. Umenweke, G.C.; Afolabi, I.C.; Epelle, E.I.; Okolie, J.A. Machine learning methods for modeling conventional and hydrothermal gasification of waste biomass: A review. *Bioresour. Technol. Rep.* **2022**, *17*, 100976. [CrossRef]
18. Xing, J.; Luo, K.; Wang, H.; Gao, Z.; Fan, J. A comprehensive study on estimating higher heating value of biomass from proximate and ultimate analysis with machine learning approaches. *Energy* **2019**, *188*, 116077. [CrossRef]
19. Afolabi, I.C.; Popoola, S.I.; Bello, O.S. Modeling pseudo-second-order kinetics of orange peel-paracetamol adsorption process using artificial neural network. *Chemom. Intell. Lab. Syst.* **2020**, *203*, 104053. [CrossRef]
20. Afolabi, I.C.; Popoola, S.I.; Bello, O.S. Machine learning approach for prediction of paracetamol adsorption efficiency on chemically modified orange peel. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2020**, *243*, 118769. [CrossRef]
21. Leng, E.; He, B.; Chen, J.; Liao, G.; Ma, Y.; Zhang, F.; Liu, S.; Jiaqiang, E. Prediction of three-phase product distribution and bio-oil heating value of biomass fast pyrolysis based on machine learning. *Energy* **2021**, *236*, 121401. [CrossRef]
22. Li, J.; Zhu, X.; Li, Y.; Tong, Y.W.; Ok, Y.S.; Wang, X. Multi-task prediction and optimization of hydrochar properties from high-moisture municipal solid waste: Application of machine learning on waste-to-resource. *J. Clean. Prod.* **2020**, *278*, 123928. [CrossRef]
23. Katongtung, T.; Onsree, T.; Tippayawong, N. Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes. *Bioresour. Technol.* **2021**, *344*, 126278. [CrossRef] [PubMed]
24. Güleç, F.; Pekaslan, D.; Williams, O.; Lester, E. Predictability of higher heating value of biomass feedstocks via proximate and ultimate analyses—A comprehensive study of artificial neural network applications. *Fuel* **2022**, *320*, 123944. [CrossRef]
25. Shenbagaraj, S.; Sharma, P.K.; Sharma, A.K.; Raghav, G.; Kota, K.B.; Ashokkumar, V. Gasification of food waste in supercritical water: An innovative synthesis gas composition prediction model based on Artificial Neural Networks. *Int. J. Hydrogen Energy* **2021**, *46*, 12739–12757. [CrossRef]
26. Elmaz, F.; Yücel, Ö.; Mutlu, A.Y. Makine Öğrenmesi ile Kısa ve Elemental Analiz Kullanarak Katı Yakıtların Üst Isı Değerinin Tahmin Edilmesi. *Int. J. Adv. Eng. Pure Sci.* **2020**, *32*, 145–151. [CrossRef]

27. Ozbas, E.E.; Aksu, D.; Ongen, A.; Aydin, M.A.; Ozcan, H.K. Hydrogen production via biomass gasification, and modeling by supervised machine learning algorithms. *Int. J. Hydrogen Energy* **2019**, *44*, 17260–17268. [CrossRef]
28. Biau, G.; Devroye, L.; Lugosi, G. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **2008**, *9*, 2015–2033.
29. Matin, S.S.; Chelgani, S.C. Estimation of coal gross calorific value based on various analyses by random forest method. *Fuel* **2016**, *177*, 274–278. [CrossRef]
30. Parikh, J.; Channiwala, S.A.; Ghosal, G.K. A correlation for calculating HHV from proximate analysis of solid fuels. *Fuel* **2005**, *84*, 487–494. [CrossRef]
31. Auret, L.; Aldrich, C. Interpretation of nonlinear relationships between process variables by use of random forests. *Miner. Eng.* **2012**, *35*, 27–42. [CrossRef]
32. Ghugare, S.B.; Tiwary, S.; Elangovan, V.; Tambe, S.S. Prediction of Higher Heating Value of Solid Biomass Fuels Using Artificial Intelligence Formalisms. *BioEnergy Res.* **2013**, *7*, 681–692. [CrossRef]
33. Tan, P.; Zhang, C.; Xia, J.; Fang, Q.Y.; Chen, G. Estimation of higher heating value of coal based on proximate analysis using support vector regression. *Fuel Process. Technol.* **2015**, *138*, 298–304. [CrossRef]
34. Çakman, G.; Gheni, S.; Ceylan, S. Prediction of higher heating value of biochars using proximate analysis by artificial neural network. *Biomass Convers. Biorefin.* **2021**, *1*, 1–9. [CrossRef]
35. Dai, Z.; Chen, Z.; Selmi, A.; Jermsittiparsert, K.; Denić, N.M.; Nešić, Z. Machine learning prediction of higher heating value of biomass. *Biomass Convers. Biorefin.* **2021**, *1*, 1–9. [CrossRef]