Journal of Engineering Science and Technology
Vol. 18, No. 1 (2023) 463 - 480

THE KRUSKAL-WALLIS TEST TO EXAMINE PERFORMANCE INCONSISTENCY IN THE BIOMASS HIGHER HEATING VALUE MODELS FOR PROXIMATE ANALYSIS

AKARA KIJKARNCHAROENSIN*, SUPACHATE INNET

School of Engineering, Unviersity of the Thai Chamber of Commerce, Thailand
*Corresponding Author: akara_kij@live4.utcc.ac.th

Abstract

Accurately determining the higher heating value (HHV) is essential for biomass to support the increasing energy demand and reduce the higher carbon dioxide levels. However, an accurate model condition on a training set might not preserve its precision in other datasets. The consistency testing method is absent. The article conducts two experiments to explore the performance consistency and test the accuracy variation of the earlier studies' HHV models. The article defined consistency as the error distribution indifference. The earlier study datasets are the treatments. The article mimicked the previous study models through a random searching algorithm on a minimum mean square error cost function. The Kruskal-Wallis H-Test determined the equivalence in the residual distributions. The results indicated the insignificant effects of the treatments on the in-sample performance at α=0.05, but the models do. The H-test can identify the accuracy inconsistency of some earlier study models in the simulated out-of-sample situations. The error distributions' non-normality limits the validity of absolute average error (AAE) and root mean square error (RMSE). A 3.4200% to 6.2923% change in AAE and a 0.7209 to 1.6117 change in RMSE do not imply the performance inconsistency of a model. The lowest AAE and RMSE are insufficient to guarantee performance consistency. Even though a model can not provide the highest accuracy on a dataset, its performance can be consistent. The Kruskal-Wallis test can successfully examine performance consistency.

Keywords: Biomass, Higher heating value, Inconsistency, Kruskal-Wallis, Prediction, Proximate analysis.

1. Introduction

Biomass is one of the most significant alternative sources of energy. The increasing energy demand and higher carbon dioxide levels during the energy crisis have adversely impacted human life [1, 2]. The subsequent emergence of new technologies has promoted biomass as a sustainable and environmentally friendly energy source [3]. Approximating the amount of heating energy, known as the higher heating value (HHV), is essential to designing an optimal process for biomass power [4]. The HHV can be estimated through two main approaches: ultimate and proximate analyses. The first approach measures biomass elements such as C, H, O, N, and S to predict the heating value, but it is a time-consuming analysis process [5]. The second approach uses amounts of fixed carbon (FC), volatile matter (VM), and ash (ASH) to evaluate the heating energy provided by biomass. Because this method is rapid and accessible [6], proximate analysis of the HHV is crucial for implementing this type of renewable energy.

Research associated with the HHV proximate analysis model continues to develop. However, some issues related to model comparison are observed in the existing literature. Previous studies [3, 7-11] evaluated their model using their dataset but compared the results with those of other already-developed models using different datasets with different characteristics. Therefore, the factors of a dataset might affect the evaluation results. Moreover, the prediction model's validation depends on the training data domain. For example, Samadi et al. [6] developed their HHV model and reported the early studies' degradation in the model accuracy [3, 7-12] on their dataset. Some literature [13-15] also said the model performs differently from the earlier studies.

There is no literature to do a formal experiment to inspect the out-of-sample accuracy of the nonlinear model. The predicted value may not be accurate when dealing with data outside the study dataset. The out-of-sample performance consistency of the model must be examined before practically implementing the HHV models.

The performance variabilities of nonlinear models are challenging to clarify owing to insufficient model information, unpublished datasets, and unrevealed source codes for reproducing the results. Xing et al. [16] compared the in-sample performance of three machine learning models on their dataset. However, these models were not the earlier study ones.

Five accuracy indicators measure the prediction error: the mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), average absolute error (AAE), and average biased error (ABE). They represent the average value of the error distributions in those mentioned studies. The average is the complete minimal sufficient statistic if the distribution is an exponential family like the normal distribution. Otherwise, it is not. The normality test is neglected in earlier studies - no recommendations for sufficient statistics in the case of non-normality.

The performance evaluations in the case of non-normality are challenging even in the same training set. The equality of these five indicators does not imply the error distribution's equivalence, while the inequality does not refer to the performance difference since the mean is not a sufficient statistic of the distribution.

This study aims to evaluate the out-of-sample performance of a proximate-based HHV model. The challengings are the measurement of the whole distribution and how to reproduce the earlier studies' models. The insufficient model information, unpublished dataset, and unrevealed source code obstructed the literature's reproduction. The normality limits the validity of present measurement indicators.

This article extends Xing et al.'s experiment to more datasets and fifteen machine learning models. These models mimic the earlier studies from random searching on the minimum MSE cost function. The highest accuracy models in each machine learner represent the earlier study ones. This study also implements two investigations and simulated out-of-sample situations to inspect the model's accuracy. The article defines accuracy consistency as the indifferent residual distributions of the prediction values. The Kruskal-Wallis H-test [17] determines the performance consistency on the null hypothesis of distribution indifference.

The article's main contributions are 1.) a statistical criterion to determine the performance consistency in the out-of-sample data points, 2.) the formal experiments to examine the model's out-of-sample performance, and 3.) the investigations of the nonlinear model performance on the out-of-sample datasets. To our knowledge, the article formally investigates the machine learning out-of-sample performance on the proximate analysis for the first time. No researcher implements the Kruska-Wallis H-test to examine the accuracy indifference on the out-of-sample data points of an HHV model.

The remainder of the article is organized as follows. Section 2 describes the experimental datasets, proxy models, and schematic experiment diagrams. The experimental results and discussions are presented in section 3. The models' out-of-sample performances and their hypothesis tests of the accuracy consistency are discussed in detail. Multiple simulated situations were conducted to statistically verify the performance consistency in some of the earlier study models. The last section is the present study's conclusion.

2. Materials and Experimental Procedures

This section describes the datasets, mimic models, and experimental designs.

2.1. Dataset from the proximate analysis

The experiments set four datasets as the treatments. The first dataset is from Qian et al. [18]. It contains 78 data points collected from [19, 20], which initially have 86 data points. Many researchers have applied these 78 data to fit their models using machine learning techniques [21-23]. The second dataset is Ghugare et al. [3]. It describes a wide range of 382 biomass data based on proximate analysis. Dashti et al. [15] use this dataset to evaluate their models. Their dataset is collected from open-source online. Nhuchhen and Afzal [8] use the third dataset. It contains 207 wide-range biomass data. Lastly, this study dataset [24] is constructed from a wide-range biomass dataset of 803 records from 17 published studies. Duplicate records are filtered out, and the older records are retained to eliminate data mismatch. The data points are on a dry basis. The unit of HHV is MJ/kg.

Figure 1 shows the entire domain of the datasets in the literature. The x, y, and z-axes refer to the amount of FC, VM, and ASH of the biomass's proximate

analysis. The amber circles refer to Qian et al.'s data [18]. Additionally, the yellow upward-pointing triangles and magenta diamond shapes refer to the datasets from Ghugare et al. [3] and Nhuchhen and Afzal [8]. Lastly, the blue plus signs represent the biomass dataset adopted in this study.

The domains of the datasets in the literature are different. They do not cover the entire FC, VM, and ASH range. For example, data from Qian et al. [18] and Nhuchhen and Afzal [8] are in the same region, a small area compared with the entire domain. The data used by Ghugare et al. [3] are categorized into small clusters and isolated from others. Because the accuracy of the training model decreases when implemented in the out-of-sample data, the trained model should be cautiously applied to the biomass, notably when the training data do not include the actual data received from that region. Therefore, in this case, the study dataset is more appropriate. The experimental datasets and source codes can be downloaded from the supplementary data linked to the study [24, 25].

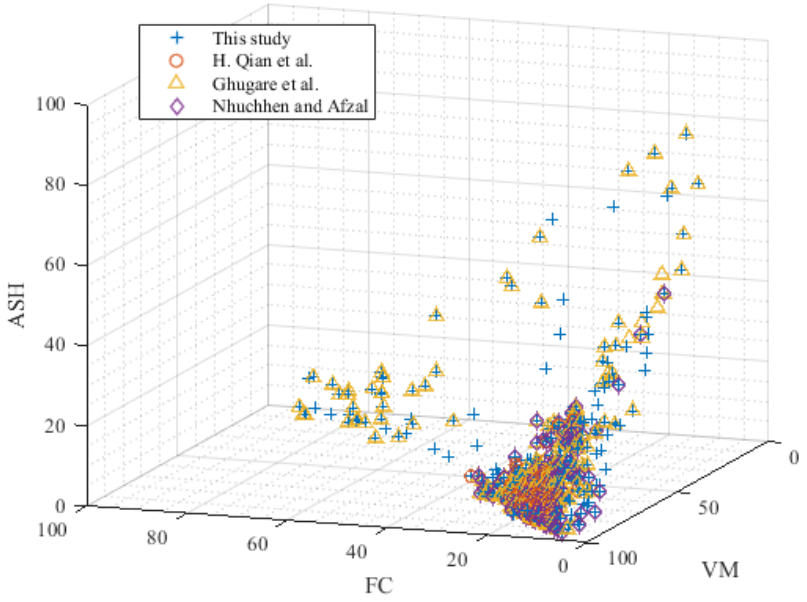

Fig. 1. Plot illustrating the datasets of the 1st and 2nd experiments.; fixed carbon (FC), volatile matter (VM), and ash (ASH).

Figure 2 shows a matrix plot of biomass data. The bar charts aligned diagonally are normalized histograms or probability density functions (pdfs). However, the distribution of the biomass data is not normal. The residuals received from the traditional linear regression method cannot be distributed normally if the features of the regression model do not have a normal distribution. Therefore, some models that require a normal distribution might be unsuitable.

The scatter plots in the upper and lower triangle parts indicate the relationship between the two parameters. The collinearities occur in VM-FC and VM-ASH feature pairs. These collinearities affect the estimated parameter variances.

Therefore, the estimated regression coefficients and variances are different and asymmetrically biased from typical values [26]. The results of the statistical hypothesis testing of the coefficients in the regression are biased.

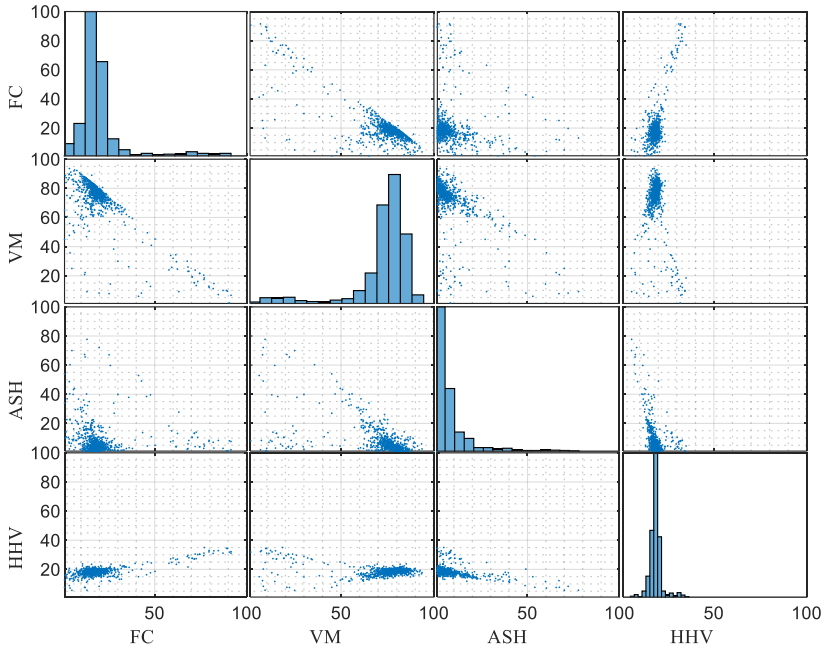

Fig. 2. Scatter plot matrix of this study biomass dataset; fixed carbon (FC), volatile matter (VM), ash (ASH), and higher heating value (HHV).

Table 1 shows the correlation matrix of biomass data. The correlation between VM and FC is -0.7842. Similarly, the correlation between VM and ASH is -0.5603. VM has a linear relationship with the FC and ASH. The estimated variances and coefficients of the linear regression model were asymmetrically biased. In contrast, FC and ASH are linearly independent.

Table 1. Correlation matrix of this study's biomass data.

Parameter	Fixed carbon (FC.)	Volatile matter (VM.)	Ash (ASH)	Higher heating value (HHV)
FC	1.0000	-0.7842	-0.0707	0.7404
VM	-0.7842	1.0000	-0.5603	-0.3008
ASH	-0.0707	-0.5603	1.0000	-0.5004
HHV	0.7404	-0.3008	-0.5004	1.0000

Figure 3 shows the overall descriptive statistics of the biomass data used in this study. The approximate range of data values is 0 to 100, and the HHV might be symmetric. Moreover, FC and ASH showed positive skewness, whereas VM exhibited negative skewness. These non-normalities affected the residual distribution, and the ordinary least squares (OLS) could not be the unique minimum variance unbiased estimator (UMVUE).

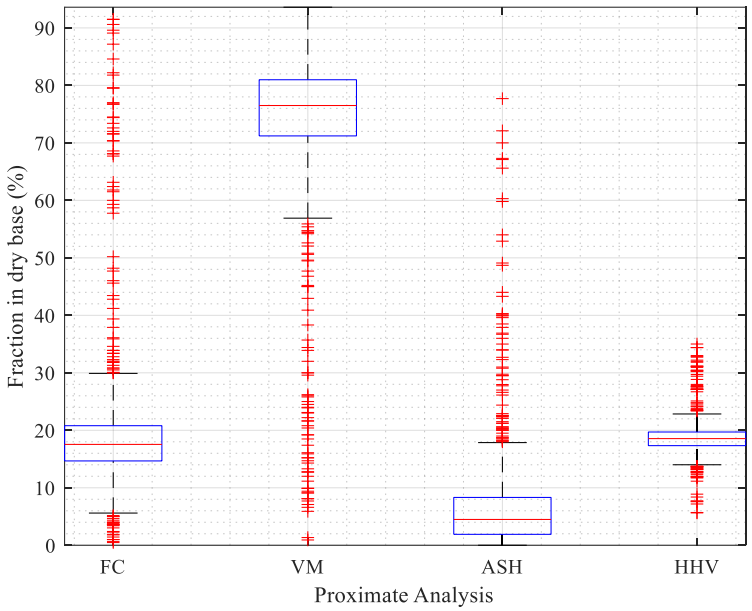

Fig. 3. Box plot of the biomass data used in this study; fixed carbon (FC), volatile matter (VM), ash (ASH), and higher heating value (HHV).

2.2. Experimental models

The article experiments implement random searching to generate fifteen mimic models of the earlier studies to test the accuracy inconsistency based on the minimum MSE cost function. MATLAB executes the random searching by setting the "OptimizeHyperparameters" equal to "all." Then, the parameters, such as basis function, kernel function, kernel scale, polynomial order, the minimum number of leaf nodes, ensemble aggregation method, and data standardization, are optimally provided to the regression learner model. This study assumes the optimization methods have minor effects on the model's accuracy. This assumption implies that the optimization method mainly contributes to the computation speed and convergence rate, not the solution vectors in the hypothesis space.

Table 2 summarizes the parameter settings for all 15 HHV mimic models. These models are 12 regression learners and three neural network models. The first six regression learner models have been studied in the literature. These include linear stepwise regression (Stepwise) [27, 28], robust linear regression (Robust) [29], support vector machine (SVM) [22], regression trees (RT) [29], ensemble trees (Ensemble) [6], and partial least squares (PLS) [30-32].

The remaining six regression learner models are as follows: the ridge regression (Ridge) [33], lasso regression (Lasso) [34], lasso generalized linear model (LassoGLM) [35], higher-dimensional (HD) linear model [36], Gaussian process (GP) regression [37] and generalized linear model (GLM) [38]. These models are applied for implementing the biomass HHV in this study for the first time. In addition, the generalized linear model can treat HHV distributions as gamma or inverse Gaussian; however, in this study, HHVs are limited to the normal distribution, and the models are limited to their quadratic terms.

The ridge, lasso, and lasso generalized linear models require the parameter lambda (λ) used in the penalty function comprising the regression learner models. The optimal λ gives the minimum sum of square residuals. This study set the λ value between 0.01 and 1000 and then searched for the optimal value, which minimizes the MSE.

Except for PLS, all linear models of regression learners are limited to the quadratic form, including constant, linear, interaction, and squared terms. The PLS model is in a linear format, not including the intercept and square terms.

The last three models are the Adaptive Neuro-Fuzzy Inference System (ANFIS), Artificial Neural Network (ANN), and Radial Basis Function (RBF). This study applied the setting reported in the literature for these models. The first model is the ANFIS-SC5 model [39]. It is an ANFIS with one 3-neuron hidden layer. This model's influence and squash factor range are 0.14 and 1.4, respectively. The second model is the ANN 13 [40], which consists of a tangent sigmoid with one 3-neuron hidden layer. The last model is GA-RBF [15], with one hidden layer and a radial basis function as the transfer function studied.

Table 2. The setting of model parameters used in the performance comparison.

	Model name	Parameter setting	Proxies
1	**Linear stepwise regression**	Quadratic	[27,28]
2	**Robust linear regression**	Quadratic	[29]
3	**Higher-dimensional linear model**	Quadratic	
4	**Generalized linear model**	Quadratic	
5	**Partial least square**	Linear	[30-32]
6	**Ridge regression**	Optimal λ, Quadratic	
7	**Lasso regression**	Optimal λ, Quadratic	
8	**Lasso generalized linear model**	Optimal λ, Quadratic	
9	**Adaptive neuro fuzzy inference system**	ANFIS-SC5	[39]
10	**Artificial neural network**	ANN 13	[40]
11	**Support vector machine**	OptimizeHyperparameters	[22]
12	**Gaussian process regression**	OptimizeHyperparameters	
13	**Ensemble tree**	OptimizeHyperparameters	[6]
14	**Regression tree**	OptimizeHyperparameters	[29]
15	**Radial basis function**	GA-RBF	[15]

2.3. The Experiments

This section describes the evaluation indicators and schematic flowcharts.

2.3.1. The measurement indicators

Five indicators represent the mimic models' accuracy. The first is the mean square error (MSE), Eq.(1), the experimental models' cost function. The second is the root mean square error (RMSE), Eq.(2), the square root of MSE. MSE and RMSE refer to both the error and its spread. They are the variance representatives. Using the MSE cost function, make Eq.(1) and Eq.(2) simultaneously represent the residual distribution's 1st and 2nd moments. Equations (3) to (5) consequently refer to the mean absolute error (MAE), average absolute error (AAE), and average biased error (ABE). MAE refers to the absolute value, while AAE and ABE refer to the relative ones. In contrast, MAE, AAE, and ABE indicate the error magnitude solely. The article evaluated the models through 20-fold cross-validation.

$$\text{MSE}=\frac{1}{N}\sum_{i=1}^{N}\left(\text{HHV}_i-\widehat{\text{HHV}_i}\right)^2 \quad (1)$$

$$RMSE=\sqrt{MSE} \tag{2}$$

$$MAE=\frac{1}{N}\sum_{i=1}^{N}\left|HHV_i-\widehat{HHV}_i\right| \tag{3}$$

$$AAE=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|HHV_i-\widehat{HHV}_i\right|}{HHV_i}x100 \tag{4}$$

$$ABE=\frac{1}{N}\sum_{i=1}^{N}\frac{\left(HHV_i-\widehat{HHV}_i\right)}{HHV_i}x100 \tag{5}$$

where N is the number of samples. HHV_i is the i^{th} sample's actual HHV while $\widehat{HHV}_i$ is its estimation. The K-fold cross-validation uses the average of K times experiment as the estimator; the more K, the narrower the interval estimation is. Twenty-fold is the maximum number implemented [23].

2.3.2. The experimental designs

The datasets are the experimental treatments to measure the accuracy variations. Two experiments define the performance inconsistency through the MSE ranking dissimilarity and residual distribution inequivalence. The experiment's null hypothesis is the model's performance consistency. The Kruskal-Wallis H test, a nonparametric ANOVA, determines the consistency through the distribution similarity.

Figure 4 exhibits the 1st and 2nd experimental diagrams—fifteen learning models involved in the experiment with min. MSE cost function. The 1st experiment's responses are the models' MSE ranking condition on the treatment. The ranking invariability is a proxy of accuracy consistency. In contrast, the 2nd experiment sets the similarity of the residual distribution as a proxy. Four datasets in the chart refer to datasets [3, 8, 18, 24]. Fifteen models are Table 2's mimic models. The moments are the functions that describe the graph's shape: the mean, variance, skewness, and kurtosis.

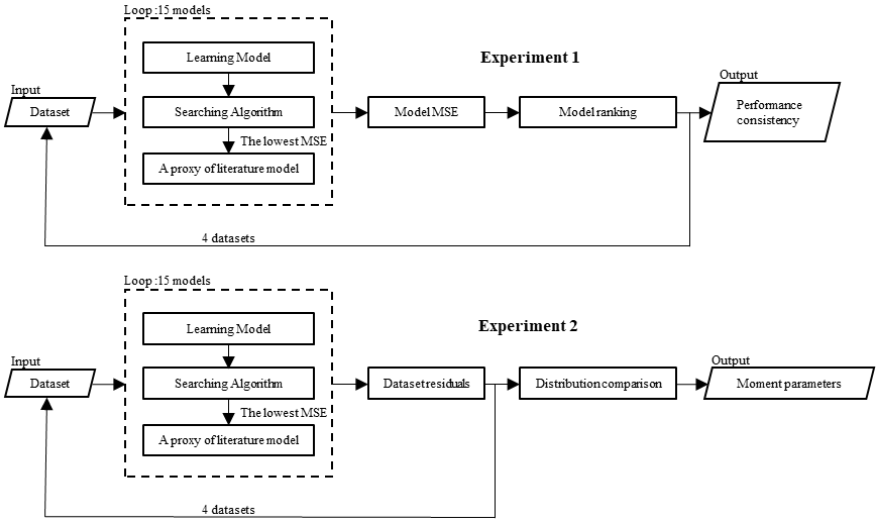

Fig. 4. Schematic flowchart to test the performance variation for the in-sample data. The datasets are the treatments. The random searching parameter constructs the fifteen proxy models based on the minimum mean square error cost function.

Let c be the number of groups, n_i be the ith group data number, and $\overline{r_i}$ be ith average rank, the Kruskal-Wallis H-test [17] statistically determines the distribution equivalence of the cumulative distributions.

$$H=\frac{12}{N(N+1)}\sum_{i=1}^{c} n_i\left(\bar{r}_i-\frac{N+1}{2}\right)^2 \quad (6)$$

Figure 5 exhibits the simulated out-of-sample situations of the HHV model implementation. The literature-reported values are treated as the in-sample performance. The out-of-sample data are the article dataset. Then, the literature models are re-evaluated on the article dataset to measure the out-of-sample performance - a total of twelve models in Table 6 (see section 3.3) involved in experiments. The H-test, Eq.(6), inspects the variation between the in and out-of-sample accuracy.

The literature implemented five measurements to assess the model performances: the RMSE, MAE, AAE (%), ABE (%), and R^2. Unfortunately, the previous studies did not report all of these measurements. The present study has to compute these measurements based on the literature's training sets to make a comparison. Then, the article alters the training set to the article dataset and re-computes these accuracy parameters.

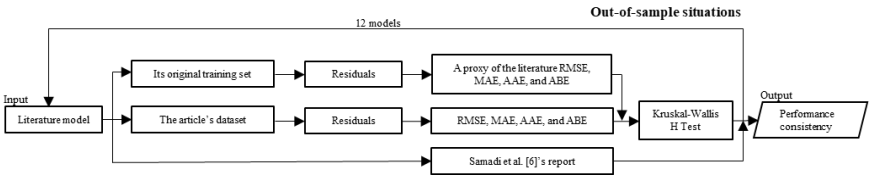

Fig. 5. Schematic flowchart of the out-of-sample simulated situations for implementing the earlier study models. Kruskal-Wallis H test on the residual distribution inspects the performance consistency.

3. Results and Discussion

The results of the 1st and 2nd experiments in Fig. 4 are reported and discussed in sections 3.1 and 3.2. The last section analyses the results of simulated out-of-sample situations.

3.1. Performance inconsistency of the model

Table 3 indicates the model's MSE and RMSE through 20-fold cross-validation. All residuals are from the testing data, which is assumed to be the out-of-sample data. The values in bold and italics are the minimum MSE of each dataset.

The magnitude of the error term was related to each treatment. The different minimum values of MSE for different models indicate that the highest accuracy model condition on a dataset does not necessarily outperform others. Moreover, MSE overshooting might occur in some datasets. This is because of the overfitting issue owing to the limited training data in the datasets. In addition, the accuracy rankings are different from [15, 41]. The radial basis network did not deliver superior accuracy to the support vector machine, artificial neural network, and adaptive neuro-fuzzy inference. However, the results support Xing et al. [16] that nonlinear models outperform linear ones. Therefore, the following investigations were conducted.

Table 3. Accuracy of biomass higher heating value (HHV) model for the four datasets based on 20-fold cross-validation.

Model	Mean square error (MSE)				Root mean square error (RMSE)			
	This study	Ghugare et al. [3]	Nhuchhen and Afzal [8]	H. Qian et al. [18]	This study	Ghugare et al. [3]	Nhuchhen and Afzal [8]	H. Qian et al. [18]
Stepwise	2.1511	1.0598	1.2521	1.0472	1.4666	1.0295	1.1190	1.0233
Robust Linear	2.1817	1.0839	1.2566	1.0825	1.4771	1.0411	1.1210	1.0404
Higher-dimensional linear	2.6306	1.3053	1.1901	***0.8672***	1.6219	1.1425	1.0909	***0.9313***
Generalized linear model (GLM)	2.1424	1.0496	1.2745	1.4288	1.4637	1.0245	1.1289	1.1953
Partial least square	2.6321	1.3039	1.2545	1.0980	1.6224	1.1419	1.1201	1.0479
Ridge	2.1309	1.0521	1.2764	1.1544	1.4598	1.0257	1.1298	1.0744
Lasso	2.1046	1.0311	1.1737	0.8683	1.4507	1.0154	1.0834	0.9318
Lasso GLM	2.5756	1.2548	***1.1657***	0.8736	1.6049	1.1202	***1.0797***	0.9347
Adaptive neuro-fuzzy inference system	2.3012	1.0931	1.4453	1671.9000	1.5170	1.0455	1.2022	40.8880
Artificial neural network	2.1140	***0.9991***	3.1161	1.9243	1.4540	***0.9996***	1.7653	1.3872
Support vector machine	2.2854	1.1467	1.2705	1.2451	1.5117	1.0709	1.1272	1.1159
Gaussian process	***2.0762***	1.0455	1.2260	1.0811	***1.4409***	1.0225	1.1073	1.0397
Ensemble	2.2283	1.3994	1.7238	1.0668	1.4927	1.1829	1.3129	1.0329
Regression tree	2.6502	1.9137	1.9553	1.4802	1.6279	1.3834	1.3983	1.2166
Radial basis function	10.3800	14.6290	3.0516	2.1999	3.2218	3.8248	1.7469	1.4832

Note: The values in bold and italics are the minimum MSE/RMSE of each dataset.

3.2. Residual distribution inconsistency

The identical distribution of the residuals is a proxy for performance consistency. The distributions comprising the dependent variable are constructed from the aggregation of the 15-model residuals obtained from the treatments. The boxplots show the distribution overview of these residuals. Subsequently, the descriptive statistics of these data are examined, and the 95% confidence intervals (CIs) of the mean errors in each dataset are determined and compared. The Kruskal-Wallis H-Test examines whether the distributions are statistically different.

Figure 6 shows the residual distributions obtained from the 15 models on every dataset, indicating that the different treatments generated different distributions. The outliers on the distribution tails refer to the probabilities that the predicted HHVs are either overestimated or underestimated. The exhibited distributions are called the fat tail in financial investment, considered a risk. Furthermore, the figure indicates that the model residuals are not normally distributed.

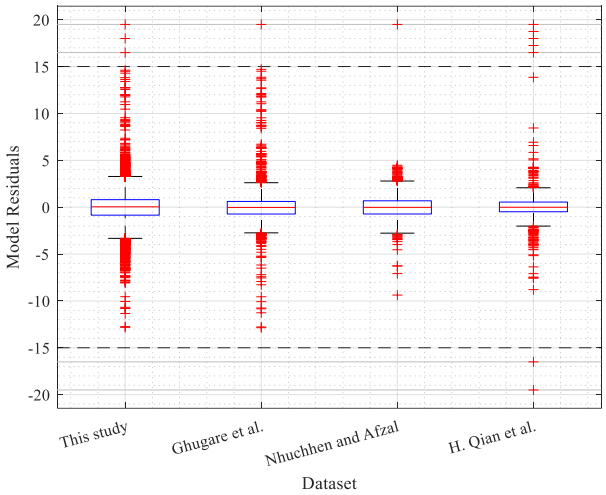

Fig. 6. Box plots of the residual distribution of all mimic models generated from four treatments.

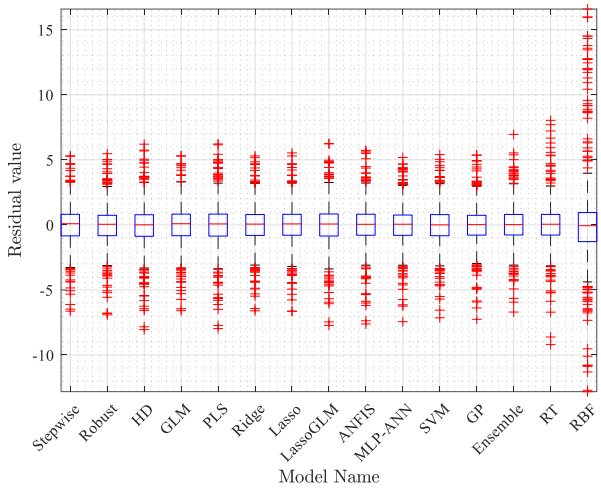

Fig. 7. Box plots of the mimic models' in-sample prediction errors on four datasets through 20-fold cross-validation.

Figure 7 expresses the 15 proxy models' in-sample prediction diversity on 1,469 data points in four datasets [3, 8, 18, 24]. The distribution distinct in Ensemble, RT, and RBF confirms the effects of models on the prediction errors, even in-sample situations. Retraining the model is insufficient unless it is consistent. Examining performance consistency is crucial for the practical implementation of the HHV models.

Table 4 presents the descriptive statistics of the residuals, including the number of observations, average, median, standard error (SE), and 95% CI of the mean. The CI indicates that the mean of the residual obtained from the Ghugare et al. [3] dataset is statistically different from zero at the 5% significance level (α=0.05). The other residuals have a zero mean at α=0.05. The statistical difference from zero of

the error average in the dataset emphasized the importance of dataset selection for training the prediction models. The study dataset provides the lowest SE (0.0154) among the four treatments.

Table 4. Descriptive statistics of the model residuals in each dataset at statistical significance, α=0.05.

Dataset	Count	Mean	Median	Skewness	Kurtosis	Standard error (SE.)	Confidence interval (CI)	
This study	12,030	-0.0027	0.0440	0.5316	12.9130	0.0154	-0.0328	0.0274
Ghugare et al. [3]	5,730	0.0389	-0.0261	2.0643	29.6070	0.0191	0.0015	0.0764
Nhuchhen and Afzal [8]	3,090	0.0241	-0.0173	1.1795	20.6950	0.0226	-0.0201	0.0684
Qian et al. [18]	1,170	-0.2088	-0.0060	-28.7730	920.8800	0.3103	-0.8177	0.4001

Table 4 reported the zero median unequal variances of the non-normal residual distributions. The study dataset generated approximately zero skewness and the lowest kurtosis among all the considered datasets. Excess kurtosis represents the probability of overestimation or underestimation. Positive skewness indicates an underestimated HHV, which leads to undersized thermal components, such as boilers and combustion chambers, and biomass overstock might occur. In contrast, the negative skewness leads to oversized specifications of the thermal process components, and the biomass feed might run out of stock.

The Kruskal-Wallis H-test in Table 5 indicates the distribution indifference even though the residual higher moments are dissimilar. The null can not be rejected at a 0.3307 p-value. The treatments are statistically insignificant to the accuracy of the proxy models. The 2nd experiment insists on the possibility that a model can preserve its accuracy among the datasets, although it can not provide the highest accuracy. The insignificant differences among the medians confirm the Kruskal-Wallis H-test results.

Table 5. The Kruskal-Wallis H-test of the model residuals in each dataset.

Source	SS.	d.f.	MS	Chi-square	p-value
Treatments	1.3839E+08	3	4.6129E+07	3.4247	0.3307
Error	8.8962E+11	22016	4.0408E+07		
Total	8.8976E+11	22019			

3.3. The out-of-sample situations

The situations of implementing the earlier study models on the out-of-sample data are illustrated in Table 6. The literature's reported performances are re-estimated on the article dataset and compared to see the performance uncertainty. The table extended Samadi et al.'s study [6] originated eight models [3,7-12] from 2012 to 2018 to eleven ones [3, 7-12, 27, 42-44] from 2001 to 2020.

The comparison results support [6] that the model performances are inconsistent among the datasets. The HHV models did not perform well in the out-of-sample data. The literature models [7, 10, 27, 42] also have a negative R^2 in the article dataset.

The Kruskal-Wallis H-test reports the null rejection at model numbers 1, 2, 3, 6, 11, and 12. RMSE and AAE are overshooting, and R^2 sign changes arise in model numbers 1 and 2 when evaluated in this study dataset. In addition, model 3 has the AAE overshooting and under-average R^2 in its training set. This improperly fitted model causes residual distribution changes when implemented in the out-of-sample data. Samadi et al. [6] also found the AAE overshooting in models 2 and 3. Models 6 and 12 have ABE overshooting issues. Samadi et al. [6] reported the AAE overshooting in model 6. They also noted the under-average R^2 of model 2 at 0.43 and that of model 6 at 0.62. In contrast, model 11 does not have problems with accuracy degradation, but the reported ABE is as high as 1.8479%. Therefore, the accuracy inconsistency arises from the improper fitting.

The H-test fails to reject the null in models 4, 5, 7, 9, and 10. In model 9, Samadi et al. [6] reported a 0.43 R^2; this study dataset encounters an R^2 sign change. Additionally, the literature said its AAE was as high as 5.880%. Model 10 has a below-average 0.4217 R^2 on its training set and a negative R^2 on this study dataset. Therefore, models 9 and 10 preserve their underperformance among the datasets. The remaining three provide their accuracy among the datasets. Thus, models 4, 5, and 7, proposed by Estiati et al. [12] and Ghugare et al. [3], provide consistent and accurate predictions in this experiment. Reallocating the original training set to the article make models 4 and 7's RMSEs change from 0.7209 to 1.6117, and 0.9410 to1.5901, while AAEs change from 3.4200% to 6.2923%, and 3.8000% to 6.6294%. Model 5's RMSE and AAE changes are similar to model 4's. The difference in accuracy indicators does not imply performance inconsistency. These simulated situations show that the Kruskal-Wallis H-test can inspect the performance consistency of the HHV model in the out-of-sample data.

4. Conclusions

The article overcame the reproduction issues of earlier studies through the mimic models generated from the random searching with min. MSE cost function. The Kruskal-Wall H-test eliminated the limitation on the normality of five accuracy indicators: MSE, RMSE, MAE, AAE, and ABE. The results insisted the change in these error averages did not imply performance inconsistency, but the H-test can statistically guarantee the performance consistency of a model among the datasets. The experiments concluded that

- The treatments have insignificant effects on the residual distributions, but the models do.
- Even though a model can not provide the highest accuracy given on a dataset, its performance can be consistent
- Error distributions' non-normality limits the validity of MSE, RMSE, MAE, AAE, and ABE. A model can preserve its accuracy between the datasets even though the RMSE and AAE change from 0.7209 to 1.6117 and 3.4200% to 6.2923%.

The article assumed the mimic model generated from random searching and min. MSE cost function can represent the earlier study model. The measurement errors in the published dataset have minor effects on the evaluations. Further research on biomass energy yield should study a factor that makes a model consistent. The accurate and consistent enhancements of the biomass HHV model support the increasing energy demand and reduce carbon dioxide levels.

Table 6. The accuracy instability of the reported HHV models due to the different datasets. The computations are the full sample resubstitution.

	Literature[I]					Samadi et al. [6] [II]			This study[III]					H-test
	RMSE	MAE	AAE(%)	ABE(%)	R^2	RMSE	AAE(%)	R^2	RMSE	MAE	AAE(%)	ABE(%)	R^2	p-value
1. Qian et al. [27]	0.7943*	0.6008*	2.5900	0.0800	0.9820				3.9907	3.0995	29.3104	25.6061	-0.5268	0.0000**
2. Qian et al.[7]	1.0518*	0.8446I	5.9800	-0.3500	0.9162	2.4100	9.1600	0.4300	3.4914	2.6303	13.3436	-7.4008	-0.1687	0.0000**
3. Nhuchhen and Afzal [8]	1.9673*	1.3800	6.1700	0.6000	0.5847*	2.1180	10.7200	0.8500	2.4219	2.0265	9.9970	-8.5861	0.4377	0.0000**
4. Estiati et al. [12]'s linear model	0.7209*	0.6048*	3.4200	0.3724*	0.9081	1.1820	4.8950	0.8800	1.6117	1.1582	6.2923	-0.4039	0.7510	0.6432
5. Estiati et al. [12]'s non-linear model	0.6885*	0.5918*	3.4700	0.4849*	0.9047	1.1210	4.6600	0.8900	1.5613	1.1220	6.1273	-0.0751	0.7663	0.7186
6. Soponpongpipat et al. [9]	0.7453*	0.5653*	2.7400	0.5200	0.9810	1.5590	8.8300	0.6200	1.8625	1.3646	7.1895	-4.0763	0.6674	0.0014**
7. Ghugare et al. [3]	0.9410	0.8623*	3.8000	3.2956*	0.9081*	1.0030	4.2230	0.9000	1.5901	1.1634	6.6294	3.0763	0.7576	0.7848
8. Muhammed et al. [11]			2.3200	-0.7600	0.9289	1.3980	5.2300	0.5800	2.6285	1.6537	22.7209	8.2613	0.3376	
9. Nuchhen and Abdul Salam [10]	1.3266*	0.9600	5.8800	0.8000	0.7918*	1.8870	5.3100	0.4300	4.7162	1.8411	10.5754	3.2241	-1.1324	0.7427
10. Yin [42]	1.3261*	1.0283*	5.4900	0.6700	0.4217*					2.2190	1.4495	7.7010	-1.0488	0.0706
11. Parikh et al. [43]	0.9010*	0.7287*	4.1933*	1.8479*	0.974*				1.6853	1.2236	6.9828	2.4101	0.7277	0.0003**
12. Cordero et al. [44]	0.5586*	0.4235*	1.5410*	-0.0070*	0.9875*				1.8202	1.3512	7.0430	-3.9280	0.6824	0.0004**

Note: [I] Data reported in the literature. [II] Samadi et al.[6] computed and reported based on their dataset. [III] Computed based on this study dataset.

* Computed in this study based on the literature's own training sets. ** Significant at $\alpha = 0.05$

[1] $HHV = -30.3FC^2 + 62.5ASH^2 + 55.4FC - 48.5ASH + 9.591$

[2] $HHV = 140.2 - 1.167FC - 0.210VM - 1.558ASH - 0.02739VM^2 + 0.000191VM^3 + 0.00104FC * ASH$

[3] $HHV = 0.1846VM + 0.3525FC$

[4] $HHV = 0.3368FC + 0.1646VM + 0.0113ASH$

[5] $HHV = 0.3241FC + 3.7667x10^{-4}FC^2 - 4.1530x10^{-4}FC * VM + 0.1947VM - 2.8207x10^{-4}VM^2 - 0.0025ASH$

[6] $HHV = 35.4879 - 0.3023ASH - 0.1905VM$

[7] $HHV = 0.365FC + 0.131VM + \frac{1.397}{FC} + \frac{328.568VM}{10283.138+0.531ASH*FC^3-6.863*ASH*FC^2}$

[8] $HHV = 1.9999 + 0.248FC + 0.162VM - 0.137ASH$

[9] $HHV = 20.7999 - 0.3214\frac{VM}{FC} + 0.0051\left(\frac{VM}{FC}\right)^2 - 11.2277\frac{ASH}{VM} + 4.4953\left(\frac{ASH}{VM}\right)^2 - 0.7223\left(\frac{ASH}{VM}\right)^3 + 0.0383\left(\frac{ASH}{VM}\right)^4 + 0.0076\frac{FC}{ASH}$

[10] $HHV = 0.2521FC + 0.1905VM$

[11] $HHV = 0.3536FC + 0.1559VM - 0.0078ASH$

[12] $HHV = 354.3FC + 170.8VM$

Nomenclatures

Greek Symbols

α	Type I error
λ	Regularization coefficients

Abbreviations

AAE	Average absolute error
ABE	Average biased error
ASH	Ash, % dry basis
CI	Confidence interval
FC	Fixed carbon, % dry basis
HHV	Higher heating value, MJ/kg
H-test	Kruskal-Wallis's test statistic
MAE	Mean absolute error
MSE	Mean square error
R^2	R-square
RMSE	Root mean square error
SE	Standard error
VM	Volatile matter, % dry basis

References

1. Keybondorian, E.; Zanbouri, H.; Bemani, A.; and Hamule, T. (2017). Application of MLP-ANN strategy to predict higher heating value of biomass in terms of proximate analysis. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 39, 2105-2111.
2. Keybondorian, E.; Zanbouri, H.; Bemani, A.; and Hamule, T. (2017). Estimation of the higher heating value of biomass using proximate analysis. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 39, 2025-2030.
3. Ghugare, S.B.; Tiwary, S.; Elangovan, V.; and Tambe, S.S. (2014). Prediction of higher heating value of solid biomass fuels using artificial intelligence formalisms. *Bioenergy Research*, 7, 681-692.
4. Ghugare, S.B.; and Tambe, S.S. (2017). Genetic programming based high performing correlations for prediction of higher heating value of coals of different ranks and from diverse geographies. *Journal of the Energy Institute*, 90, 476-484.
5. Akkaya, A.V. (2013). Predicting coal heating values using proximate analysis via a neural network approach. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 35, 253-260.
6. Samadi, S.H.; Ghobadian, B.; and Nosrati, M. (2021). Prediction of higher heating value of biomass materials based on proximate analysis using gradient boosted regression trees method. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 43, 672-681.
7. Qian, X.; Lee, S.; Soto, A.-m.; and Chen, G. (2018). Regression model to predict the higher heating value of poultry waste from proximate analysis. *Resources*, 7(3), 39.

8. Nhuchhen, D.R.; and Afzal, M.T. (2017). HHV predicting correlations for torrefied biomass using proximate and ultimate analyses. *Bioengineering*, 4(1).
9. Soponpongpipat, N.; Sittikul, D.; and Sae-Ueng, U. (2015). Higher heating value prediction of torrefaction char produced from non-woody biomass. *Frontiers in Energy*, 9, 461-471.
10. Nhuchhen, D.R.; and Abdul Salam, P. (2012). Estimation of higher heating value of biomass from proximate analysis: A new approach. *Fuel*, 99, 55-63.
11. Mohammed, I.Y.; Kazi, F.K.; Suzana, Y.; Alshareef, I.M.; and Chin, S.A. (2014). Higher heating values (HHV) prediction model from biomass proximate analysis data. *Conference: International Conference & Exhibition on Clean Energy (ICCE 2014)*, Quebec, Canada, 292-300.
12. Estiati, I.; Freire, F.B.; Freire, J.T.; Aguado, R.; and Olazar, M. (2016). Fitting performance of artificial neural networks and empirical correlations to estimate higher heating values of biomass. *Fuel*, 180, 377-383.
13. Hosseinpour, S.; Aghbashlo, M.; Tabatabaei, M.; and Mehrpooya, M. (2017). Estimation of biomass higher heating value (HHV) based on the proximate analysis by using iterative neural network-adapted partial least squares (INNPLS). *Energy*, 138, 473-479.
14. Ceylan, Z.; Pekel, E.; Ceylan, S.; and Bulkan, S. (2018). Biomass higher heating value prediction analysis by ANFIS, PSO-ANFIS and GA-ANFIS. *Global Nest Journal*, 20, 589-597.
15. Dashti, A.; Noushabadi, A.S.; Raji, M.; Razmi, A.; Ceylan, S.; and Mohammadi, A.H. (2019). Estimation of biomass higher heating value (HHV) based on the proximate analysis: Smart modeling and correlation. *Fuel*, 257, 115931.
16. Xing, J.; Luo, K.; Wang, H.; Gao, Z.; and Fan, J. (2019). A comprehensive study on estimating higher heating value of biomass from proximate and ultimate analysis with machine learning approaches. *Energy*, 188, 116077.
17. Kruskal, W.H. (1952). A nonparametric test for the several sample. *The Annals of Mathematical Statistics*, 23, 525-540.
18. Qian, H.; Guo, X.; Fan, S. et al. (2016). A simple prediction model for higher heat value of biomass. *Journal of Chemical and Engineering Data*, 61, 4039-4045.
19. Song, G.; Shen, L.; and Xiao, J. (2011). Estimating specific chemical exergy of biomass from basic analysis data. *Industrial and Engineering Chemistry Research*, 50, 9758-9766.
20. Vassilev, S.V.; Baxter, D.; Andersen, L.K.; and Vassileva, C.G. (2010). An overview of the chemical composition of biomass. *Fuel*, 89, 913-933.
21. Darvishan, A.; Bakhshi, H.; Madadkhani, M.; Mir, M.; and Bemani, A. (2018). Application of MLP-ANN as a novel predictive method for prediction of the higher heating value of biomass in terms of ultimate analysis. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 40, 2960-2966.
22. Duan, M.; Liu, Z.; Yan, D.; Peng, W.; and Baghban, A. (2018). Application of LSSVM algorithm for estimating higher heating value of biomass based on ultimate analysis. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 40, 709-715.
23. Ighalo, J.O.; Adeniyi, A.G.; and Marques, G. (2020). Application of linear regression algorithm and stochastic gradient descent in a machine-learning

environment for predicting biomass higher heating value. *Biofuels, Bioproducts and Biorefining*, 14, 1286-1295.

24. Kijkarncharoensin, A. (2022). Proximate_analysis. Mendeley Data. Retrieved January 30, 2022, from https://doi.org/10.17632/g36dhg826s.2.
25. Kijkarncharoensin, A. (2022). Performance inconsistency of the biomass higher heating value (HHV) models derived from proximate analysis. Codeocean. Retrieved January 25, 2022, from https://doi.org/10.24433/CO.0255376.v1.
26. Mela, C.F.; and Kopalle, P.K. (2002). The impact of collinearity on regression analysis: The asymmetric effect of negative and positive correlations. *Applied Economics*, 34, 667-677.
27. Qian, C.; Li, Q.; Zhang, Z.; Wang, X.; Hu, J.; and Cao, W. (2020). Prediction of higher heating values of biochar from proximate and ultimate analysis. *Fuel*, 265, 116925.
28. Telmo, C.; Lousada, J.; and Moreira, N. (2010). Proximate analysis, backwards stepwise regression between gross calorific value, ultimate and chemical analysis of wood. *Bioresource Technology*, 101, 3808-3815.
29. Vallejo, F.; Díaz-Robles, L.A.; Vega, R.; and Cubillos, F. (2020). A novel approach for prediction of mass yield and higher calorific value of hydrothermal carbonization by a robust multilinear model and regression trees. *Journal of the Energy Institute*, 93, 1755-1762.
30. Posom, J.; Shrestha, A.; Saechua, W.; and Sirisomboon, P. (2016). Rapid non-destructive evaluation of moisture content and higher heating value of Leucaena leucocephala pellets using near infrared spectroscopy. *Energy*, 107, 464-472.
31. Posom, J.; and Sirisomboon, P. (2017). Evaluation of lower heating value and elemental composition of bamboo using near infrared spectroscopy. *Energy*, 121, 147-158.
32. Posom, J.; and Sirisomboon, P. (2017). Evaluation of the higher heating value, volatile matter, fixed carbon and ash content of ground bamboo using near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, 25, 301-310.
33. Hoerl, A.E.; and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.
34. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267-288.
35. Friedman, J.; Hastie, T.; and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1-22.
36. Ho, C.H.; and Lin, C.J. (2012). Large-scale linear support vector regression. *Journal of Machine Learning Research*, 13, 3323-3348.
37. Bousquet, O.; Luxburg, U. von; and Rätsch (eds.), G. (2003). *Advanced lectures on machine learning*. ML Summer Schools 2003. Berlin : Springer, c2004.
38. Zheng, B.; and Agresti, A. (2000). Summarizing the predictive power of a generalized linear model. *Statistics in Medicine*, 19 (2000), 1771-1781.
39. Akkaya, E. (2016). ANFIS based prediction model for biomass heating value using proximate analysis components. *Fuel*, 180, 687-693.

40. Uzun, H.; Yıldız, Z.; Goldfarb, J.L.; and Ceylan, S. (2017). Improved prediction of higher heating value of biomass using an artificial neural network model based on proximate analysis. *Bioresource Technology*, 234, 122-130.
41. Taki, M.; and Rohani, A. (2022). Machine learning models for prediction the higher heating value (HHV) of municipal solid waste (MSW) for waste-to-energy evaluation. *Case Studies in Thermal Engineering*, 31, 101823.
42. Yin, C.Y. (2011). Prediction of higher heating values of biomass from proximate and ultimate analyses. *Fuel*, 90,1128-1132.
43. Parikh, J.; Channiwala, S.A.; and Ghosal, G.K. (2005). A correlation for calculating HHV from proximate analysis of solid fuels. *Fuel*, 84, 487-494.
44. Cordero, T.; Marquez, F.; Rodriguez-Mirasol, J.; and Rodriguez, J. (2001). Predicting heating values of lignocellulosics and carbonaceous materials from proximate analysis. *Fuel*, 80, 1567-1571.