Quantitative Assessment Of Operational Components Regarding Supercritical Water Gasification Systems' Hydrogen Yield

Yu Qian

Fushun Vocational Technology Institute, Fushun 113122, Liaoning, China

Corresponding author. E-mail: 13130353333@163.com

Received: December 31, 2025; Accepted: May 04, 2026

Supercritical Water Gasification is a promising technology for the production of hydrogen-rich gas from biomass, but it suffers from high operating costs because of the extreme temperature and pressure requirements. A reasonable estimation for hydrogen produced by this process is required for the optimization of process and scaling it for industrial purposes. The present study aims to provide a predictive framework for hydrogen mole fraction estimation in the absence of expensive experimental trials. For accurate modelling of water gasification high-dimensional space, the current research provides comparative machine learning based farmwork, which are optimized using a novel Pufferfish Optimization Algorithm by hyperparameter optimization strategy. The performance comparison indicates that optimized Extremely Gradient Boosting model achieves the highest performance in hydrogen mole fraction prediction by yielding Coefficient of Determination of 0.9992, Mean Squared Error of 0.2325 during testing process. The interpretability analysis of this model demonstrates that the Temperature and Catalyst Type parameters influential parameters, that higher catalyst type negatively impact the model output, while elevated temperatures positively influence the model output.

Keywords: Supercritical water gasification; Hydrogen; Quantitative assessment; Wet biomass conversion

© The Author('s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.

http://dx.doi.org/10.6180/jase.202701_36.006

1. Introduction

Within the limiting global temperature constraints rise (to 1.5 °C), the population increase and the industrial revolution pose serious challenges regarding energy sustainability by potentially raising its consumption [1]. Climate change and fossil fuel depletion are two major threats to the future viability of our species. Fossil fuels dominate the contemporary energy landscape, accounting for approximately 84% of the global energy supply [2]. For instance, combustion of fossil fuels is responsible for the emission of greenhouse gas, which is held accountable for global warming, shifting the earth's average temperature, and hence climate change, natural resource depletion, and erratic weather patterns. Thus, it has harmful direct and indirect impacts on the environment and ecosystems [3]. In this regard, there is a pressing need for Renewable Energy Sources (RES) enhancement [4]. The global pursuit of sustainable RES has intensified interest in hydrogen (H_2) as a prominent alternative to fossil fuels, considering its high energy content and zero-emission profile, making it suitable for various applications, including transportation, power generation, and industrial processes [5]. One of the most effective methods for H_2production is the Supercritical Water Gasification (SCWG) of biomass and organic industrial wastes [6]. This thermochemical process converts wet biomass materials, such as agricultural residues, food waste, and industrial sludge, into H_2 rich synthesis gas under supercritical water conditions [7]. Operating at temperatures above 374°C and high pressures, SCWG enables rapid and efficient gasification while avoiding the energy intensive drying step required in conventional gasification processes [8]. As a

result, SCWG improves efficiency of energy and offers a promising pathway for sustainable H_2 production from waste-based resources.

The SCWG process's complexity arises from various parameters, including biomass composition, reaction temperature, pressure, residence time, and catalyst type. These factors significantly influence the hydrogen mole fraction (H_2 mole fraction) and other composition in the product gas. Furthermore, identifying the optimal operating conditions through experimental trials alone is challenging, as SCWG experiments require extremely high temperatures and pressures, making them time consuming and costly. To address this limitation, modeling approaches have been expanded to predict process performance more efficiently [9]. In this context, accurate prediction of H_2 mole fraction is essential for optimizing SCWG operations and facilitating industrial-scale implementation [10, 11]. Machine Learning (ML) offers the benefit of speed and accuracy in calculations. These models can find hidden trends and patterns through the utilization of high-dimensional data and sophisticated computing methods, without requiring complex rule-based algorithms [12]. Furthermore, these algorithms permit real-world scenarios to significantly lower computational costs compared to common mechanistic models without assumptions. They are trained with the initial dataset and predict results based on data from subsequent or other inputs [13].

Recent advancements have seen the integration of ML techniques to model and predict outcomes of the SCWG process. For instance, a Generative Adversarial Network (GAN) for data augmentation, coupled with Least Squares Support Vector Regression (LSSVR), to predict H_2 mole fraction, achieving a test Coefficient of Determination $\left(R^2_{\text{test}}\right)$ of 0.9457 [14]. A comparative study of four ML models to economically predict H_2 mole fraction, from biomass during the SCWG process. Their analysis evaluated the predictive capability of several algorithms and demonstrated that the Random Forest (RF) model delivered the highest accuracy by achieving an R^2_{test} at 0.9782 [15]. Proposed a hybrid ML-thermo-dynamic framework for predicting H_2 mole fraction from the SCWG process. Their approach integrated Artificial Neural Network (ANN) with a modified Gibbs-free energy minimization method, which achieved the reasonable R^2_{test} of 0.92 [16]. On the other hand, optimization frameworks, which aimed at improving the predictive capability and generalization performance of ML models, have emerged as promising approaches. In several studies, metaheuristic-based optimization techniques have been employed to identify the optimal set of hyperparameters for ML models [17]. To achieve this, hyperparameter optimization techniques across multiple algorithms, such as Extreme Gradient Boosting (XGB) and RF for H_2 mole fraction prediction [18]. Similarly, the models as ANN, Ensembled Learning Tree (ELT), Support Vector Machine (SVM), Gaussian Process Regression (GPR), with Particle Swarm Optimization (PSO) to predict H_2 mole fraction from SCWG process. According to the results, ELT-PSO was preferred as the best H_2 mole fraction predictor model, obtaining an R^2_{test} of 0.997 [19]. Furthermore, an ML-based framework incorporating Decision Trees (DT), Gradient Boosting Regression (GBR), ELT and SVM, which were further enhanced using Differential Evolution (DE). Among these models, the ELT-DE approach achieved the best performance with an R^2_{test} of 0.950 [20]. Several ML models optimized using PSO and Genetic Algorithms (GA) based on 12 process and biomass features derived from 166 data points. Among the models tested, boosting algorithms, particularly XGB demonstrated superior performance by achieving an R^2_{test} value of 0.84 for H_2 mole fraction prediction [21]. In another article, eight ML models to predict hydrogen H_2 mole fraction from biomass's SCWG using a dataset with 237 samples. Their results indicated that a GBR model optimized with PSO achieved the highest robustness, with an R^2_{test} of 0.958 [22].

Despite the increasing use of ML methods to predict H_2 mole fraction from biomass during SCWG, several important limitations remain in existing studies. (i) Many existing studies rely on linear or relatively simple models, which may not be capable of capturing the complex nonlinear relationships inherent in SCWG processes. As a result, these approaches may fail to fully exploit the underlying patterns present in experimental data. (i) Some studies depend on default or manually selected hyperparameters designs to simplify model development and speed up predictions. The neglecting optimization of hyperparameter can significantly reduce model's predictive performance, particularly when dealing with the SCWG systems with dynamics environment. (iii) Many investigations do not incorporate model interpretability or feature importance analyses, using methods such as Shapley Additive Explanations (SHAP) techniques, which are essential for understanding the contribution and influence of input variables on model predictions.

Given the recognized advantages of ML-based model optimization, along with the growing need for interpretable predictive frameworks for the SCWG process of biomass, this study was designed to address several critical research gaps through a comprehensive modeling and analysis approach. (i) The study utilizes the comparative modelling approach using advanced ML algorithms, namely Adap-

tive Boosting (AdaBoost), Extremely Randomized Trees (ET), Bayesian Ridge Regression (BRR), Huber Regression (HR), SVM, and XGB to predicts H_2 mole fraction form SCWG process. (ii) It ensures optimal predictive capability and robustness of the developed models by implementing a systematic hyperparameter optimization strategy using a novel Pufferfish Optimization Algorithm (POA), which enhances the accuracy, and stability of each ML model. (iii) This work sought to improve the interpretability and mechanistic understanding of the SCWG process by applying sensitivity and feature importance analysis using the SHAP and permutation importance method, which enables for detailed evaluation of how individual process variables influence H_2 mole fraction predictions.

In section 2, the study approach and the performance metrics used, information about the research dataset, and arrangements applied to the information are elaborated. The results are analyzed, evaluated, and represented by a variety of analytical charts to measure the performance of different classifiers based on assessment indicators in Section 3. Finally, in section 4, overall findings are summarized by the study's conclusion.

2. Methodology

The methodology aims to improve the SCWG procedure's data-driven understanding with a focus on H_2 mole fraction prediction by using some preprocessing steps, advanced modeling, optimization, and interpretability studies. As an effort, Fig. 1 visualizes the procedures for this prediction purposes. The workflow starts with collecting the experimental dataset related to the SCWG process of biomass, which includes the input variables describing biomass properties and operating conditions, along with the target variable, which is the H_2 mole fraction to be predicted. The dataset is prepared before model training using complete preprocessing methods, which cleans and prepares the dataset for developing process. Then, the processed dataset is divided into Training data and Test data in which the K-fold cross-validation (CV) are used instead of a simple split to improve reliability of model evaluation. The XGB, SVM, ET, AdaBoost, BRR, and HR algorithms are selected as candidate models for predicting H_2 mole fraction and trained using train dataset. During training process, for each selected model, the most suitable hyperparameter configuration is defined using a population-based POA optimization algorithm. Therefore, the models are trained using these parameters, evaluated on the validation data, and their performance (fitness) are calculated using evaluation criteria. To interpret the best model (XGB)'s predictions, the SHAP analysis and permutation importance are applied, which identify the most influential input features affecting the prediction of H_2 mole fraction.

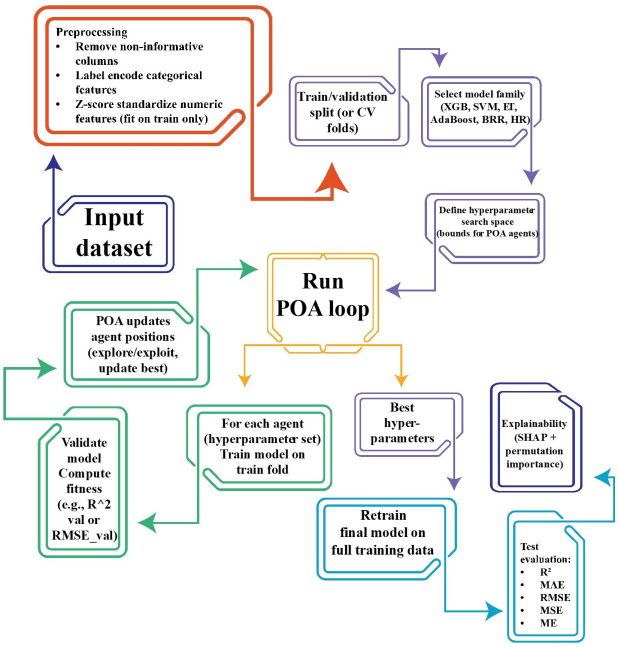


Fig. 1. The procedure of developing ML models for H_2 mole fraction predictions form SCWG process of biomass.

2.1. Data Collecting and Preprocessing

2.1.1. *Dataset description*

The dataset gathered from [20] includes a total of 244 samples obtained from SCWG experiments conducted on biomass (polymer wastes). Each data point includes complete input parameters such as temperature, pressure, residence time, feedstock type, and catalyst composition, along with several output gas variables, which are detailed in Table 1. Each dataset sample records the feedstock type and amount along with its chemical composition, including elemental contents (C, H, N, O, S) and ash. It also contains indicators of key functional groups and important SCWG operating conditions. The process outputs syngas with composition of mole fractions of Carbon dioxide (CO_2), H_2, Carbon monoxide (CO), methane (CH_4) and H_2 selectivity. This study selects the H^2 mole fraction as the target variable.

2.1.2. *Dataset preprocessing*

Before training the ML model, the dataset undergoes a series of following preprocessing steps:

- First, non informative or irrelevant columns such as Reactor that do not contribute meaningful information to the prediction task are removed from the dataset.

Table 1. Detailed information about the SCWG dataset's features and target

Feature	Definition	Range (min–max)	Unit	Categories
Feed	Feedstock identity	0–7	–	8
Type of Feed	Feedstock class/type	0–23	–	24
C, H, N, O, S	Ultimate analysis composition	C: 25.7–60.8; H: 4.2–8.7; N: 0–11.2; S: 0–2.8; O: 26.1–56.6	wt.%	–
ASH	Ash content	0–25	wt.%	–
C − H Aromatics	Aromatic C − H indicator	0–1	(0/1)	–
C − H Alkanes	Aliphatic C − H indicator	0–1	(0/1)	–
C − O	C − O bond indicator	0–1	(0/1)	–
C = C	C = C bond indicator	0–1	(0/1)	–
C = O	C = O bond indicator	0–1	(0/1)	–
O − H	O − H bond indicator	0–1	(0/1)	–
Temp	Reaction temperature	80–400	°C	–
Time	Residence time	1–240	min	–
BC	Catalyst concentration/loading	4–50	(as reported)	–
Catalyst Type	Catalyst identity	0–15	–	16
Target: H_2 mole fraction	H_2 in syngas	0–65	mol%	–

- Second, the remaining data is carefully prepared for numerical processing. Since many ML algorithms require numerical inputs, categorical variables such as Feed, Type of Feed and Catalyst Type are transformed into numerical representations using label encoding method. In this process, each unique category within a categorical feature is assigned a distinct integer value, allowing the algorithm to process these variables while preserving their categorical distinctions.

- Third, numerical features are standardized using Z score normalization, which results in features with a mean of zero and a standard deviation of one. Importantly, to prevents data leakage, the parameters required for standardization are calculated exclusively from the training dataset.

2.1.3. *Data splitting and K-fold CV*

From a total dataset of 244 samples, an 80/20 stratified split is employed to ensure robust model development and evaluation. Specifically, 194 samples (80%) are allocated to for model fitting, while the remaining 50 samples (20%) are reserved as an independent test set for final performance assessment. The generalization capability of the six developed ML models is evaluated using Repeated 5-fold CV prior to hyperparameter optimization. To prevent overfitting or leakage, a Repeated 5-fold CV strategy is applied within the training phase. This process involves partitioning the 194 training samples into five distinct folds that in each iteration, the model was trained on four folds and validated on the remaining one-fold. This 5-fold process was repeated multiple times with different random shuffles to stabilize the performance estimates and ensure that the model's predictive accuracy is not a result of a specific favorable data split.

2.2. Machine Learning Model Training

As is obvious in Fig. 1, to handle the intricacies of nonlinear, high-dimensional, noisy data, six unique ML models were chosen. The models are provided by sources in the following.

2.2.1. *Extremely Randomized Tree*

ET and RF are members of the ensemble learning algorithm class. To complete a task, ensemble learning algorithms combine the strengths of learning algorithms to improve the final prediction. ET, construct multiple trees like RF algorithms at the training phase on the entire dataset but on different feature sets. It is important to mention that although bootstrapping is not used in ET's initial form, it could be used in certain variants [23]. Furthermore, when constructing every DT, the ET algorithm splits nodes randomly. With varying subsets of data, a range of biases can be achieved in the result; however, ET works against this by taking samples from the entire dataset. Another benefit of ET is that it can decrease variance. This is because of the randomization employed in the node splitting in the decision trees; hence, the algorithm is less susceptible to some features or patterns in the data [24].

2.2.2. *Support Vector Machine*

The SVM technique is applied to both classification and regression problems. For solving these issues, it employs a kernel function to connect the input and output mathematically. The function is quite useful in complicated situations where nonlinearities exist; the mathematical function is

provided fully in [25]. Kernel functions can be defined in SVM decision-making, and Python provides the option of custom kernels. This approach can deal with linear classification and regression problems without needing major changes in hyperparameters [26].

2.2.3. *Bayesian Ridge Regression*

BRR employs Bayesian inference for the derivation of the probability distribution of the model parameters, making it different from the normal linear regression. Classical linear regression provides single, deterministic parameter values, but BRR considers them in a probabilistic manner as uncertain quantities [27]. BRR is a useful tool for calculating hydrogen generation yields from biomass, even if it may not be used as frequently in this field as other models, like RF or SVM. This is because it can manage multicollinearity and offer uncertainty estimates [28]. For example, regarding the structure of biomass, temperature, pressure, residence time, merging expert knowledge, and accounting for uncertainty, BRR produces more robust and transparent models, particularly useful for biomass conversion datasets, which are complex and volatile [29]. The learning procedure of BRR model is detailed in Fig. 2.

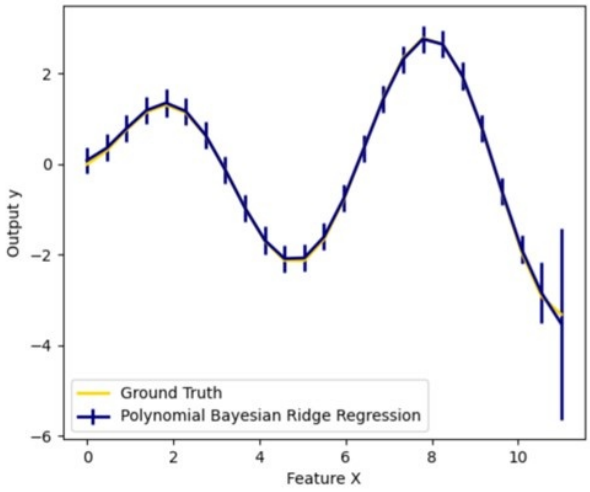


Fig. 2. The learning procedure of BRR model.

2.2.4. *Huber Regression*

The application of the Huber loss function to ML models, especially for regression problems, is referred to as the HR. This can be derived by combining two extremely popular loss functions, such as MSE and MAE [30]. Huber loss is an actual ML method for handling noisy data and outliers by taking advantage of the strengths of MSE and MAE. For large errors, it is similar to MAE (absolute loss) and decreases the dependency of the model on outliers [31]. Fig. 3 demonstrates the HR's learning procedure.

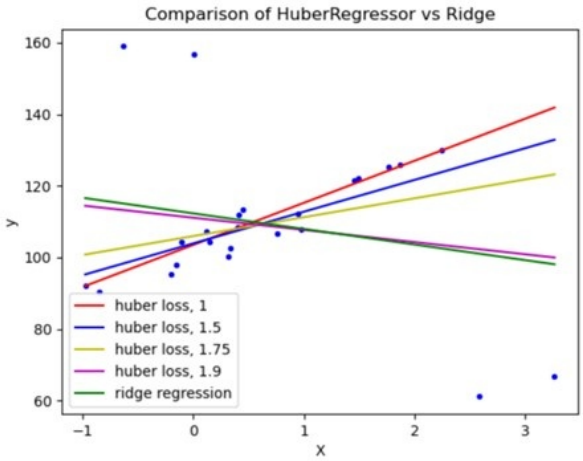


Fig. 3. The learning procedure of HR model.

2.2.5. *Extreme Gradient Boosting*

A boosting algorithm that is efficient and scalable and uses second-order gradients and regularization to regulate overfitting. This algorithm can be applied effectively for structured or tabulated data processing. XGB is a very effective and well-optimized gradient boosting technique. It improves the model's accuracy by using boosted and normalizing approaches [32]. There are high levels of viability of the XGB algorithm in H_2 mole fraction prediction of biomass conversion because of its robustness and ability to handle complicated nonlinearities [33].

2.3. Hyperparameter Tuning Using POA

The attribute values of hyperparameters have significant effects on both the learning and predictive accuracies of models. The research uses the POA to make the process of hyperparameter tuning for each model. The first population consisted of 10 agents, and random configurations of hyperparameters were chosen. Then the POA completed 100 repetitions, where the agent positions are refined by the R^2 score on the validation data to determine fitness. Using principles of evolutionary approaches, the algorithm attempts to find the balance between exploration of a more general parameter landscape and improvement of the found optimal candidate regions. Finally, different sets of hyperparameters were tuned for each model. Pufferfish Optimization Algorithm

The POA imitates pufferfish defensive behavior and is a bio-inspired metaheuristic optimization method. The method functions in two principal modes, concluding exploration and exploitation [34]. Thus, the exploration phase encourages the search agents to examine the entire area for

answers by simulating how the pufferfish might react to approaching predators. The exploitation phase replicates the pufferfish's inflating defense, directing the search agents to focus more intently on areas where better answers are probable to be located. The location of each search agent in the POA is updated mathematically by (1) [34]:

The attribute values of hyperparameters have significant effects on both the learning and predictive accuracies of models. The research uses the POA to make the process of hyperparameter tuning for each model. The first population consisted of 10 agents, and random configurations of hyperparameters were chosen. Then the POA completed 100 repetitions, where the agent positions are refined by the R^2 score on the validation data to determine fitness. Using principles of evolutionary approaches, the algorithm attempts to find the balance between exploration of a more general parameter landscape and improvement of the found optimal candidate regions. Finally, different sets of hyperparameters were tuned for each model.

Pufferfish optimization algorithm

The POA imitates pufferfish defensive behavior and is a bio-inspired metaheuristic optimization method. The method functions in two principal modes, concluding exploration and exploitation [34]. Thus, the exploration phase encourages the search agents to examine the entire area for answers by simulating how the pufferfish might react to approaching predators. The exploitation phase replicates the pufferfish's inflating defense, directing the search agents to focus more intently on areas where better answers are probable to be located. The location of each search agent in the POA is updated mathematically by (1) [34]:

$$X_i^{t+1} = X_i^t + r.\left(X_{\text{best}}^t - X_i^t\right) + \alpha.\delta \qquad (1)$$

Where in (1): X_i^t is the position of the agent i at iteration t.

X_{best}^t is the best-known position at iteration t.

r is a random number in the range [0,1].

α is a scaling factor controlling the influence of the random component.

δ is a random vector representing the stochastic nature of the search.

The POA has been effectively applied in various optimization problems, including parameter tuning in ML models and engineering design tasks. In the context of H_2 production from SCWG of biomass, the POA can be employed to optimize process parameters such as temperature, pressure, and feedstock composition to maximize H_2 mole fraction. This approach enhances the efficiency and sustainability of hydrogen production processes.

As an example, Fig. 4 provides the process for XGB hyperparameters, including nestimators, learning rate, and max-depth to optimize. The details from each iteration, the bestperforming configurations for all models, were saved in the Optimization folder.

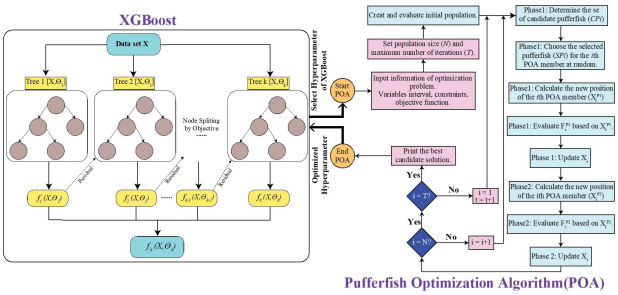


Fig. 4. Representation of the XGB's optimization process by the POA metaheuristic model.

2.4. Model Evaluation Metrics

The performance of models was measured with these metrics, as provided in Table 2, in which y_i and $\hat{y}_i$ refer to the actual and predicted values, while n indicates number of samples. R^2 determines the ability to explain variability in the target variables due to the input variables. MAE serves as a straightforward average quantification of the absolute discrepancy between predicted and actual values; hence, it is easy to explain how the model performs in real situations. RMSE barely focuses on significantly huge errors, offering superior evidence of the model's likely outliers. ME shows the biggest mistake between predicted and actual values, which is an important indication of substantial model malfunctions. MSE estimates the mean square error between the predictions that your model makes and the actual values, or the sum of the squares of the errors.

Table 2. The metric formula for the prediction of machine learning.

Criteria	Equation
Mean Absolute Error	$MAE = \frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}{n}$
Root Mean Squared Error	$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$
Max Error	$ME = \max\left(\lvert y_i - \hat{y}_i\rvert\right)$
Mean Square Error	$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$
Coefficient Determination	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - 1/n(\sum_{i=1}^{n} y_i))^2}$

Table 3. The results from Repeated 5-fold CV for six developed ML models before applying optimization.

Model	R^2(CV)	RMSE (CV)	MSE (CV)	MAE (CV)	ME (CV)
SVM	0.9921 ± 0.004	1.215 ± 0.18	1.476 ± 0.41	0.624 ± 0.09	5.98 ± 1.42
XGB	0.9987 ± 0.001	0.521 ± 0.06	0.272 ± 0.07	0.305 ± 0.05	2.41 ± 0.52
BRR	0.9153 ± 0.021	4.862 ± 0.54	23.64 ± 5.10	3.57 ± 0.42	15.88 ± 2.34
HR	0.9038 ± 0.026	5.104 ± 0.63	26.05 ± 6.01	3.69 ± 0.47	17.44 ± 2.71
AdaBoost	0.9975 ± 0.002	0.612 ± 0.09	0.374 ± 0.11	0.392 ± 0.07	2.02 ± 0.61
ET	0.9979 ± 0.002	0.598 ± 0.08	0.357 ± 0.10	0.276 ± 0.06	3.54 ± 0.88

3. Results and discussion

3.1. Repeated 5-fold CV Results

The generalization capability of the six developed ML models is depicted using the results of repeated 5-fold CV in Table 3. As summarized in the results, the XGB model demonstrated superior performance, by achieving the highest R^2 of 0.9987 ± 0.001 and the greatest level of stability across folds. Furthermore, the exceptionally small error metrics observed for the XGB model underscores its high robustness and reliability, which established it as the most effective baseline for further optimization and application in this study. This top-tier performance was followed by the ET and AdaBoost ensemble models, which also showed strong predictive accuracy by registering R^2 of 0.9979 ± 0.002 and 0.9975 ± 0.002. In contrast, linear-based models, specifically BRR and HR, yielded significantly lower R^2 values and higher error dispersion. This performance gap indicates that linear frameworks possess a limited capacity

3.2. Performance Comparison Results

The obtained results from evaluation of six optimized models in H_2 mole fraction prediction form SCWG process is depicted in Table 4. According to that, the optimized-XGB model demonstrated outstanding predictive performance compared with the other evaluated models. During the training phase, XGB achieved extremely high accuracy with an R^2 of 0.9999 , which indicating that almost all variance in the target variable was explained by the model. The error metrics were also very low, which reflected minimal deviation between predicted and actual values. Importantly, the model maintained strong performance on the testing dataset, where it achieved an R^2 of 0.9992 , along with RMSE of 0.4821 , MSE of 0.2325 , and MAE of 0.2923 . The relatively small difference between training and testing errors indicates that the XGB model generalizes well and does not suffer from significant overfitting. This superior performance of XGB was because it effectively captured non-linear relationships and highorder feature interactions inherent in the SCWG dataset using gradient boosting framework, which sequentially minimized residual errors, and led to refined predictive accuracy. Additionally, built-in regularization (L1/L2), shrinkage (learning-rate), and subsampling reduce variance and improve generalization compared to single learners (such as HR, and BRR) and less-regularized ensembles (such as AdaBoost, and ET). Therefore, XGB provides an optimal bias-variance balance, explaining its lower test MSE and higher R^2 compared to the other models despite relatively small sample size and heterogeneous feature space.

Beyond the top-performing XGB model, AdaBoost emerged as the second-most effective predictor, demonstrating exceptional stability with an R^2 of 0.9989 and a competitive RMSE of 0.5655 during testing. In contrast, both SVM and ET exhibited a lower efficiency especially during testing, which indicates their limitation in this specific prediction task. On the lower end of the spectrum, the simpler models of BRR and HR were the weakest performers so that their test RMSE values were approximately ten times higher than leading method.

Table 4. The results from evaluation of six optimized models.

	R^2	MAE	MSE	RMSE	ME
Train					
SVM	0.9999	0.0225	0.0048	0.0695	0.4623
XGB	0.9999	0.0635	0.0099	0.0995	0.6246
BRR	0.9476	2.8065	13.4989	3.6741	19.3001
HR	0.9473	2.6741	13.5810	3.6852	21.6702
AdaBoost	0.9991	0.3414	0.2308	0.4805	2.1979
ET	0.9999	0.0657	0.0144	0.1200	0.9354
Test					
SVM	0.9932	0.6860	1.9955	1.4126	6.7831
XGB	0.9992	0.2923	0.2325	0.4821	2.2220
BRR	0.9218	3.6520	22.7939	4.7743	14.5457
HR	0.9085	3.7486	26.6947	5.1667	16.7037
AdaBoost	0.9989	0.4171	0.3197	0.5655	1.5879
ET	0.9987	0.2599	0.3920	0.6261	3.9213

For more proof of the results, Fig. 5 compares the obtained results for evaluation metrics visually. According to that, XGB registered the most favorable balance between accurate prediction and minimizing errors. The XGB model indicating superior prediction capability by achieving R^2 of at 0.99 during testing and achieved lowest error rates, which reflected minimal deviation between predicted and

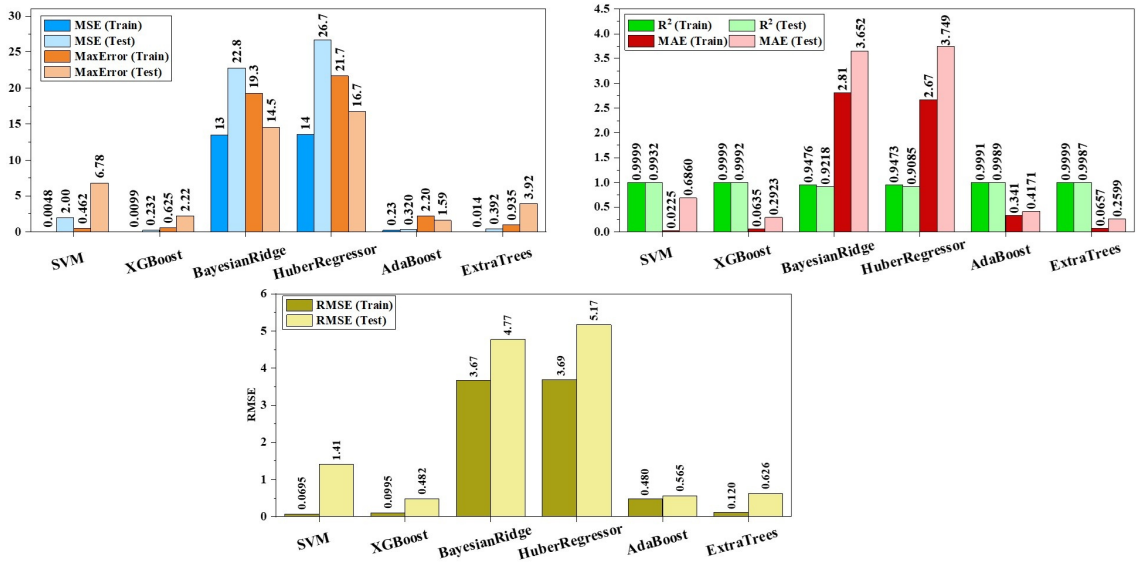


Fig. 5. The ML visual comparisons using obtained results from the performance metrics.

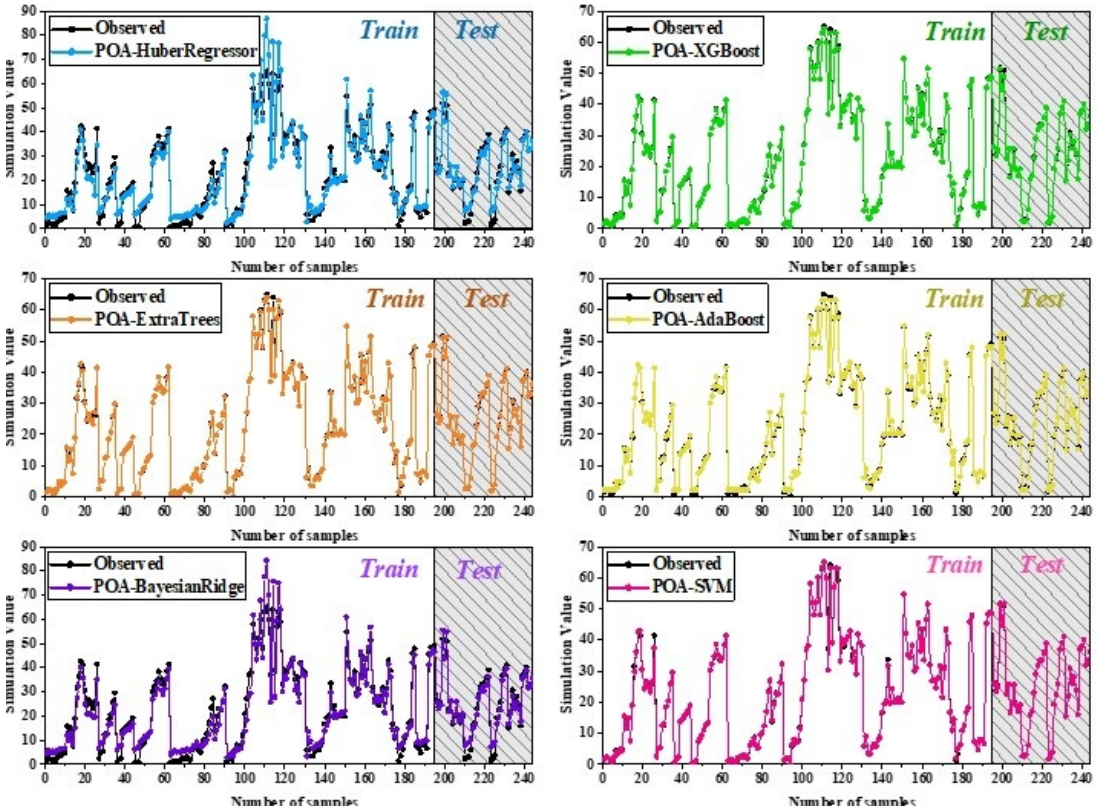


Fig. 6. The alignment level between predictions by each optimized ML model and actual values over data samples.

actual values. The AdaBoost also indicates the comparable results, while SVM and ET during test indicates that recognizing non-linear relationships compared to the XGB model were weaker. The BRR and HR suffer a much higher error and a lesser R^2 on the test data, which shows that they do not provide proper consideration of the influence of key features. The underperformance of these models is probably due to their limited ability to capture nonlinearity and feature interaction.

Fig. 6 provides critical information on model's predic-

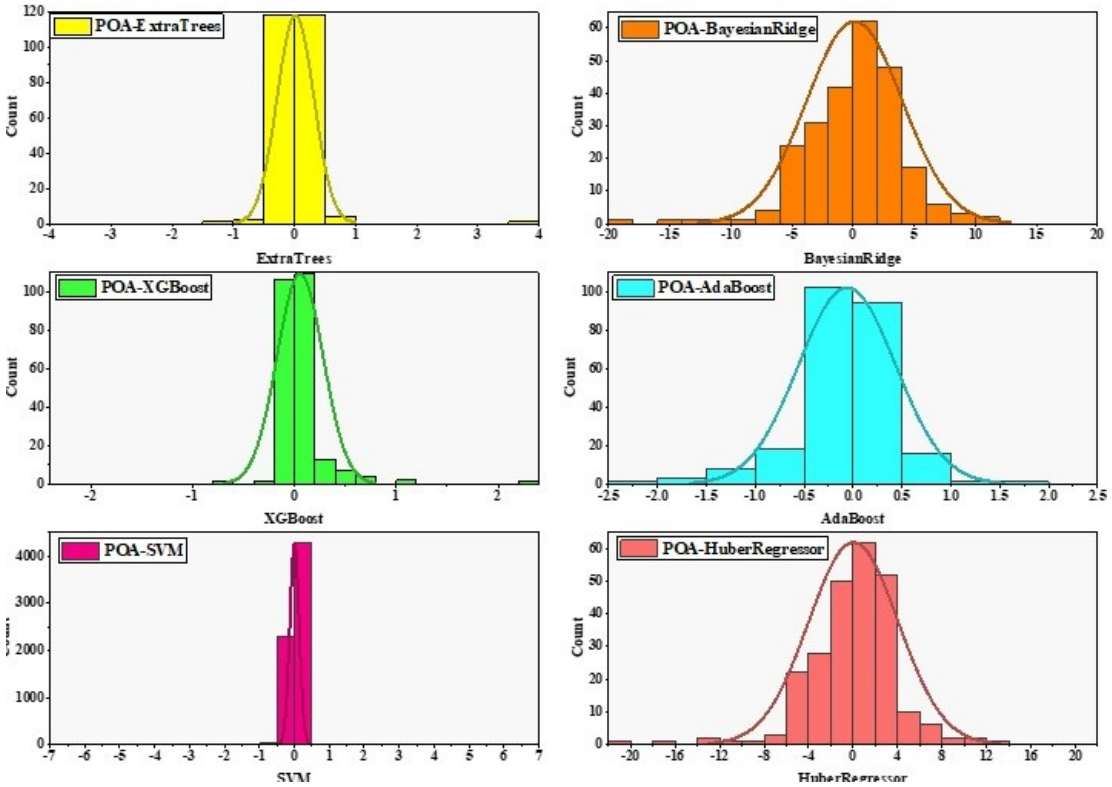
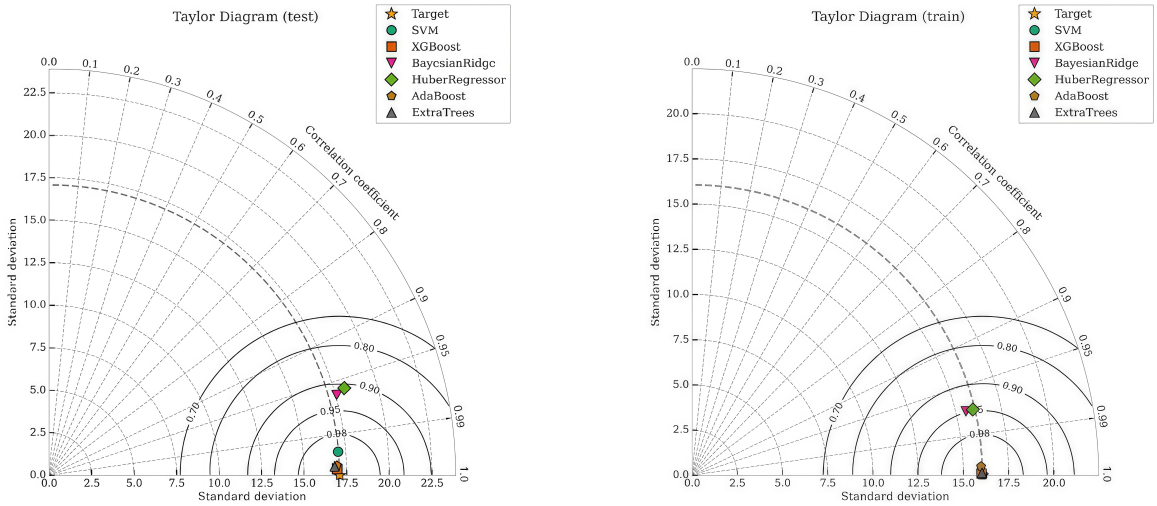


Fig. 7. The error distribution trend comparison for six optimized models.



(a) Train **(b)** Test

Fig. 8. Taylor diagram comparison for optimized six model during train and test.

tion skills through the alignment level between simulation data (predictions) and actual values (observed) over data samples. XGB proved to be the most capable model in comparison to all models. In training as well as testing, the prediction curve demonstrates a high degree of accordance of predictions with actual observations. Such consistent performance validates the hypothesis that the model holds excellent abilities to generalize to the data, indicating a balance between bias and variance. The small and centralized residuals of XGB further attest to its domain stability and suggest strongly controlled bias and variance. AdaBoost took a conservative approach in mak-

ing its prediction. While having an acceptable global error rate, the model does not possess the ability to show all the intricacies of the target signal. The SVM and TE models, however, saw its predictive performance plummet when it shifted to the test domain. Though the model performed satisfactorily during the training period by predicting comparatively accurate results, deviations from such accurate predictions were noted when the model was tested on the test set, demonstrating significant and consistent deviations from actual results. In contrast, HR and BBR through the abandonment of utilizing tracking for every minute change, were sometimes unable to control the intricacies in high-frequency data. In general, these models were capable of handling moderate deviations optimally, but its performance is affected if it encounters highly non-linear dynamics or relatively fast-changing dynamics.

The analysis of the error distribution trend in Fig. 7 offers additional visual clarity regarding the findings. The results show that models, ensemble-based models, particularly XGB, and AdaBoost exhibit the most favorable error patterns, characterized by higher concentration of errors around zero and normal distribution patterns. The XGB demonstrated the most balanced error distribution and maintained minimal deviation between predictions and actual values, due to its ability for handling dynamic and non-linear relations within dataset. ET and SVM also performs well, but slightly distribution of errors implies the presence of a few larger deviations. Meanwhile, BRR and HR present the largest error magnitudes and the widest distribution of error in broader range, likely due to their limited capacity to capture the nonlinear relationships in the dataset.

The Taylor diagram in Fig. 8 explains the performance of models by integrating three key metrics of correlation coefficient, R^2 and the standard deviation, into a single plot. In the context of the evaluated models, a Taylor diagram visually confirmes the superiority of ensemble methods such as XGB, AdaBoost and ET by positioning them closest to the Reference point (the actual observed data). This proximity indicates that these models not only achieved a high correlation (near 1.0) but also accurately replicated the variability (standard deviation) of the original dataset. Conversely, the BRR and HR models were further from the reference point, which highlights their lower correlation and higher error margins.

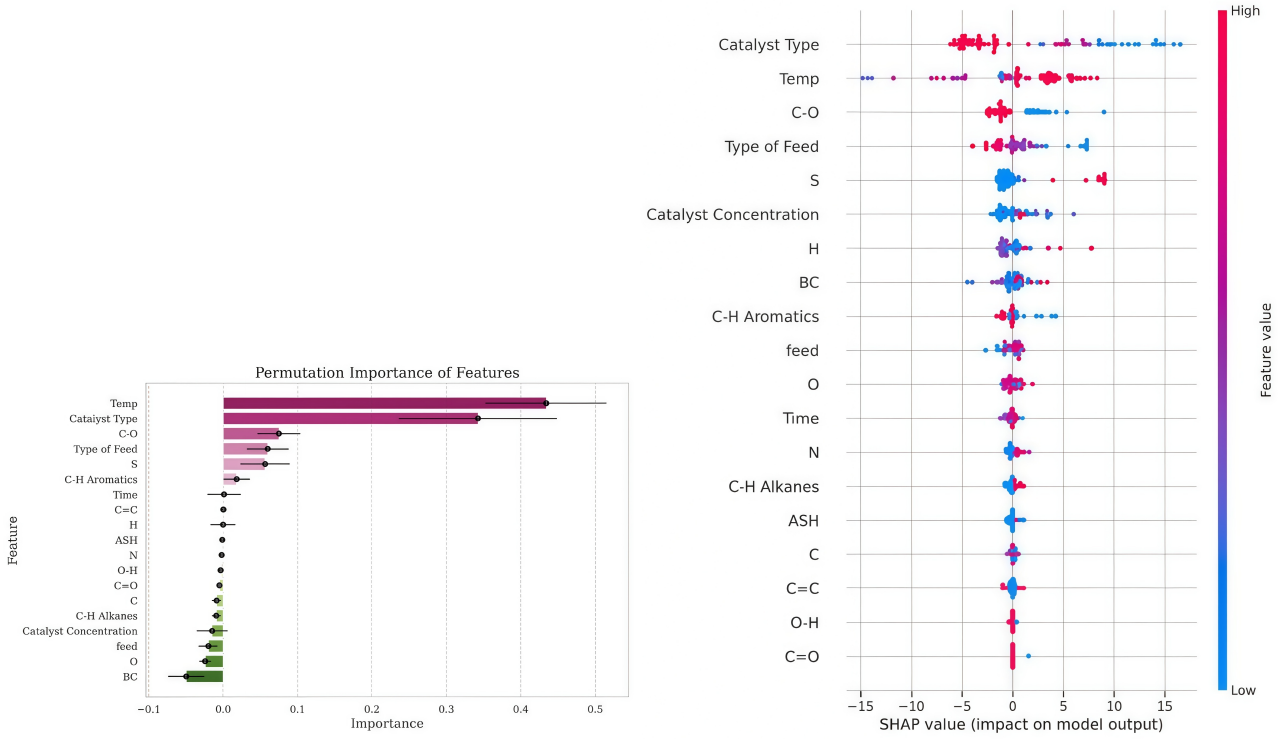
3.3. XGB's Interpretability Analysis Results

The results of interpretability analysis of XGB model in prediction of H_2 mole fraction prediction form SCWG process by applying sensitivity and feature importance analysis using the SHAP and permutation importance method is depicted in Fig. 9(a and b). Foremost among the influential parameters are Temp and Catalyst Type. The analysis reveals that the Catalyst Type holds the highest importance score, approximating 4.5, closely followed by Temp with a score of around 3.5. Delving deeper into their individual impacts, the SHAP sensitivity analysis indicates a nuanced relationship for Catalyst Type so that higher values of the catalyst type, likely corresponding to a specific catalyst variety, negatively impact the XGB model output and, consequently, the H_2 mole fraction. Conversely, lower values for this parameter demonstrate a positive effect on H_2 mole fraction. In contrast, the Temp parameter exhibits a more direct correlation so that elevated temperatures positively influence the XGB model output and H_2 mole fraction, while lower temperatures contribute negatively to the SCWG process and the overall H_2 mole fraction. Beyond these dominant factors, other parameters such as the C–O bond indicator, Type of Feed, and S content also play significant roles (importance score between 0.05 to 0.1). Specifically, high values for the C–O bond indicator and Type of feed are found to negatively contribute to the XGB model output and H_2 mole fraction. Conversely, a higher S content demonstrates a positive contribution to the model's output and the H_2 mole fraction, which suggests that an increased sulfur content can surprisingly enhance the efficiency of the SCWG process, leading to a greater output of hydrogen.

4. Conclusion

The research sought to improve the supercritical water gasification procedure using data-driven modelling with a focus on H_2 mole fraction prediction. The study utilized the comparative modelling approach using advanced Adaptive Boosting, Extremely Randomized Trees, Bayesian Ridge Regression, Huber Regression, Support Vector Machine, and Extremely Gradient Boosting. It ensured optimal predictive capability and robustness of the developed models by implementing a systematic hyperparameter optimization strategy using a novel Pufferfish Optimization Algorithm. Furthermore, this work sought to improve the interpretability of develop model by applying sensitivity and feature importance analysis.

The performance comparison revealed that ensemble-based models, particularly Extremely Gradient Boosting optimized with Pufferfish Optimization Algorithm achieved the highest performance, yielding R^2 0.9992, along with RMSE of 0.4821, MSE of 0.2325, and MAE of 0.2923. This model showed superior performance because it effectively captured complex non linear relationships and



(a) Permutation importance

(b) SHAP sensitivity

Fig. 9. XGB's interpretability analysis results using Permutation importance and SHAP sensitivity in prediction of H_2 mole fraction prediction form SCWG process.

feature interactions in the dataset using a gradient boosting framework that sequentially minimized prediction errors. Its built in mechanisms, regularization, learning rate shrinkage, subsampling, and hyperparameter optimization helped to reduce variance and prevent overfitting. For this reason, Extremely Gradient Boosting rocked up as the best model, which highlights predictive accuracy to generalization, and clear interpretability.

The interpretability analysis of Extremely Gradient Boosting model indicates that the foremost among the influential parameters were Temperature and Catalyst Type parameters. The analysis reveals that the Catalyst Type holds the highest importance score, that with higher values negatively impact the model output and, consequently, the H_2 mole fraction. In contrast, the Temperature parameter exhibits a more direct correlation so that elevated temperatures positively influence the model output and H_2 mole fraction. Therefore, these findings indicate that the proposed strategy enables a more informed, data driven, and faster optimization of the gasification process by clearly identifying the most influential parameters and clarifying their directional impacts on H_2 mole fraction.

Funding

This work was supported by the National Natural Science Foundation of China (grant number 52106208), and Central Government Guided Local Science and Technology Development Fund Project (2024ZYD0123).

nomenclature

Acronyms

AdaBoost	Adaptive Boost
ANN	Artificial Neural Network
BRR	Bayesian Ridge Regressor
DT	Decision Tree algorithm
H2	Hydrogen
MAE	Mean Absolute Error
ME	Max Error
ML	Machine Learning
POA	Pufferfish Optimization Algorithm
RES	Renewable Energy Source
RF	Random Forest
RMSE	Root Mean Square Error
SCWG	Supercritical Water Gasification
SHAP	Shapley Additive explanations
SVM	Support Vector Machine
SVR	Support Vector Regression
XGB	Extreme Gradient Boost

References

[1] M. Tauseef, A. A. Ansari, A. H. Khoja, S. R. Naqvi, R. Liaquat, W. Nimmo, and S. S. Daood, (2022) *"Thermokinetics synergistic effects on co-pyrolysis of coal and rice husk blends for bioenergy production"* **Fuel *318***: 123685. DOI: https://doi.org/10.1016/j.fuel.2022.123685.

[2] F. F. S. Supply. *Percent of World Energy—And Other Eye Openers from BP's Annual Review*. 84.

[3] M. A. Azni, R. M. Khalid, U. A. Hasran, and S. K. Kamarudin, (2023) *"Review of the effects of fossil fuels and the need for a hydrogen fuel cell policy in Malaysia"* **Sustainability *15***: 4033. DOI: https://doi.org/10.3390/su15054033.

[4] M. A. H. Khan, S. Bonifacio, J. Clowes, A. Foulds, R. Holland, J. C. Matthews, C. J. Percival, and D. E. Shallcross, (2021) *"Investigation of biofuel as a potential renewable energy source"* **Atmosphere *12***: 1289. DOI: https://doi.org/10.3390/atmos12101289.

[5] M. M. H. Bhuiyan and Z. Siddique, (2025) *"Hydrogen as an alternative fuel: A comprehensive review of challenges and opportunities in production, storage, and transportation"* **International Journal of Hydrogen Energy *102***: 1026–1044. DOI: https://doi.org/10.1016/j.ijhydene.2025.01.033.

[6] A. Kruse, (2008) *"Supercritical water gasification"* **Biofuels, Bioproducts and Biorefining: Innovation for a sustainable economy *2***: 415–437. DOI: https://doi.org/10.1002/bbb.93.

[7] G. Ruoppolo, P. Ammendola, R. Chirone, and F. Miccio, (2012) *"H2-rich syngas production by fluidized bed gasification of biomass and plastic fuel"* **Waste Management *32***: 724–732. DOI: https://doi.org/10.1016/j.wasman.2011.12.004.

[8] J. A. Okolie, S. Nanda, A. K. Dalai, F. Berruti, and J. A. Kozinski, (2020) *"A review on subcritical and supercritical water gasification of biogenic, polymeric and petroleum wastes to hydrogen-rich synthesis gas"* **Renewable and sustainable energy reviews *119***: 109546. DOI: https://doi.org/10.1016/j.rser.2019.109546.

[9] T. L. Saaty and M. Sodenkamp, (2008) *"Making decisions in hierarchic and network systems"* **International Journal of Applied Decision Sciences *1***: 24–79. DOI: https://doi.org/10.1504/IJADS.2008.017952.

[10] C. Wang, C. Hu, Y. Zheng, H. Jin, and Z. Wu, (2023) *"Predictive control of reactor network model using machine learning for hydrogen-rich gas and biochar polygeneration by biomass waste gasification in supercritical water"* **Energy *282***: 128441. DOI: https://doi.org/10.1016/j.energy.2023.128441.

[11] O. Yakaboylu, J. Harinck, K. G. Smit, and W. D. Jong, (2015) *"Supercritical water gasification of biomass: a literature and technology overview"* **Energies *8***: 859–894. DOI: https://doi.org/10.3390/en8020859.

[12] S. K. Bishnu, S. Y. Alnouri, and D. M. Al-Mohannadi, (2023) *"Computational applications using data driven modeling in process Systems: A review"* **Digital chemical engineering *8***: 100111. DOI: https://doi.org/10.1016/j.dche.2023.100111.

[13] M. Marković, M. Bossart, and B.-M. Hodge, (2023) *"Machine learning for modern power distribution systems: Progress and perspectives"* **Journal of Renewable and Sustainable Energy *15***: 032301. DOI: https://doi.org/10.1063/5.0147592.

[14] Z. Ma, J. Wang, Y. Feng, R. Wang, Z. Zhao, and H. Chen, (2023) *"Hydrogen yield prediction for supercritical water gasification based on generative adversarial network data augmentation"* **Applied Energy *336***: 120814. DOI: https://doi.org/10.1016/j.apenergy.2023.120814.

[15] W. Zhao, Y. Chen, J. Wu, Z. Zhu, and J. Wang, (2021) *"Study on CO2 capture upgrading of existing coal fired power plants with gas steam mixture cycle and supercritical water coal gasification"* **International Journal of Greenhouse Gas Control *112***: 103482. DOI: https://doi.org/10.1016/j.ijggc.2021.103482.

[16] M. Mohamadi-Baghmolaei, P. Zahedizadeh, M. Khajeh, S. Zendehboudi, and N. M. C. Saady, (2024) *"Developing a hybridized thermodynamic and data-driven model for catalytic supercritical water gasification of biomass for hydrogen production"* **Energy Conversion and Management *307***: 118302. DOI: https://doi.org/10.1016/j.enconman.2024.118302.

[17] Z. U. Haq, H. Ullah, M. N. A. Khan, S. R. Naqvi, and M. Ahsan, (2022) *"Hydrogen production optimization from sewage sludge supercritical gasification process using machine learning methods integrated with genetic algorithm"* **Chemical Engineering Research and Design *184***: 614–626. DOI: https://doi.org/10.1016/j.cherd.2022.06.020.

[18] C. C. Ukwuoma, D. Cai, A. L. Jonathan, N. Chen, C. Sey, N. W. Ntia, O. Bamisile, and Q. Huang, (2024) *"Enhancing hydrogen production prediction from biomass gasification via data augmentation and explainable AI: A comparative analysis"* **International Journal of Hydrogen Energy *68***: 755–776. DOI: https://doi.org/10.1016/j.ijhydene.2024.04.283.

[19] M. N. A. Khan, Z. U. Haq, H. Ullah, S. R. Naqvi, U. Ahmed, M. Zaman, and N. A. S. Amin, (2024) *"Prediction of hydrogen yield from supercritical gasification process of sewage sludge using machine learning and particle swarm hybrid strategy"* **International Journal of Hydrogen Energy *54***: 512–525. DOI: https://doi.org/10.1016/j.ijhydene.2023.01.033.

[20] S. Azadvar and O. Tavakoli, (2024) *"Data-driven interpretation, comparison and optimization of hydrogen production from supercritical water gasification of biomass and polymer waste: Applying ensemble and differential evolution in machine learning algorithms"* **International Journal of Hydrogen Energy *85***: 511–525. DOI: https://doi.org/10.1016/j.ijhydene.2024.08.081.

[21] K. Khandelwal and A. K. Dalai, (2024) *"Prediction of individual gas yields of supercritical water gasification of lignocellulosic biomass by machine learning models"* **Molecules *29***: 2337. DOI: https://doi.org/10.3390/molecules29102337.

[22] K. Khandelwal, S. Nanda, and A. K. Dalai, (2025) *"Machine learning modeling of supercritical water gasification for predictive hydrogen production from waste biomass"* **Biomass and Bioenergy *197***: 107816. DOI: https://doi.org/10.1016/j.biombioe.2025.107816.

[23] S. Heddam. "Intelligent data analytics approaches for predicting dissolved oxygen concentration in river: extremely randomized tree versus random forest, MLPNN and MLR". In: Springer, 2020, 89–107. DOI: https://doi.org/10.1007/978-981-15-5772-9_5.

[24] F. Afshar, S. Seyedabrishami, and S. Moridpour, (2022) *"Application of Extremely Randomised Trees for exploring influential factors on variant crash severity data"* **Scientific reports *12***: 11476. DOI: https://doi.org/10.1038/s41598-022-15693-7.

[25] R. Al-Qasimi and F. Al-Hajri, (2024) *"The Implementation of a Support Vector Regression Model Utilizing Meta-Heuristic Algorithms for Predicting Undrained Shear Strength"* **Advances in Engineering and Intelligence Systems *3***: 49–66. DOI: https://doi.org/10.22034/aeis.2024.489661.1253.

[26] H. Alkattan, B. Turyasingura, B. Willbroad, and A. A. K. Jaafar, (2025) *"Economic performance classification in Iraq (2000–2023): A statistical analysis using machine learning with support vector machines and random forest"* **EDRAAK *2025***: 29–37. DOI: https://doi.org/10.70470/EDRAAK/2025/005.

[27] M. Imane, E.-S. Aoula, and E. H. Achouyab. "Using Bayesian ridge regression to predict the overall equipment effectiveness performance". In: *2022 2nd International conference on innovative research in applied science, engineering and technology (IRASET)*. IEEE, 2022, 1–4. DOI: https://doi.org/10.1109/IRASET52964.2022.9738316.

[28] E. Y. Kaya, I. Ali, Z. Ceylan, and S. Ceylan, (2024) *"Prediction of higher heating value of hydrochars using Bayesian optimization tuned Gaussian process regression based on biomass characteristics and process conditions"* **Biomass and Bioenergy *180***: 106993. DOI: https://doi.org/10.1016/j.biombioe.2023.106993.

[29] S. Rodgers, A. Bowler, L. Wells, C. S. Lee, M. Hayes, S. Poulston, E. Lester, F. Meng, J. McKechnie, and A. Conradie, (2024) *"A surrogate model for the economic evaluation of renewable hydrogen production from biomass feedstocks via supercritical water gasification"* **International Journal of Hydrogen Energy *49***: 277–294. DOI: https://doi.org/10.1016/j.ijhydene.2023.08.016.

[30] Y. Feng and Q. Wu, (2022) *"A statistical learning assessment of Huber regression"* **Journal of Approximation Theory *273***: 105660. DOI: https://doi.org/10.1016/j.jat.2021.105660.

[31] M. K. Çodur, (2023) *"Ensemble machine learning approaches for prediction of Türkiye's energy demand"* **Energies *17***: 74. DOI: https://doi.org/10.3390/en17010074.

[32] R. Sibindi, R. W. Mwangi, and A. G. Waititu, (2023) *"A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices"* **Engineering Reports *5***: e12599. DOI: https://doi.org/10.1002/eng2.12599.

[33] S. Zhao, J. Li, C. Chen, B. Yan, J. Tao, and G. Chen, (2021) *"Interpretable machine learning for predicting and evaluating hydrogen production via supercritical water gasification of biomass"* **Journal of Cleaner Production *316***: 128244. DOI: https://doi.org/10.1016/j.jclepro.2021.128244.

[34] O. Al-Baik, S. Alomari, O. Alssayed, S. Gochhait, I. Leonova, U. Dutta, O. P. Malik, Z. Montazeri, and M. Dehghani, (2024) *"Pufferfish optimization algorithm: a new bio-inspired metaheuristic algorithm for solving optimization problems"* **Biomimetics *9***: 65. DOI: https://doi.org/10.3390/biomimetics9020065.