



A Machine Learning Framework for Accurate Prediction of Biohydrogen Production from Dark Fermentation of Food Wastes

Nour Elislam MOUGARI ^{a,c,*}, Djamel Eddine GHERSI ^b, Farida Iachachene ^{a,c}, Hanane CHERADI ^a

^a *Laboratory of Energy, Mechanics and Engineering (LEMI), Department of Mechanical Engineering, Faculty of Technology, University M'Hamed Bougara Boumerdes (UMBB), Frantz Fanon city Boumerdes 35000, Algeria*

^b *Laboratory of Theoretical and Applied Fluid Mechanics, Faculty of Physics, University of Science and Technology Houari Boumediene, BP 32 Bab Ezzouar, 16111, Algiers, Algeria.*

^c *Earth Physics Laboratory, Faculty of Hydrocarbon and Chemistry, University M'Hamed Bougara Boumerdes, Boumerdes 35000, Algeria*

*Corresponding Author Email: n.mougari@univ-boumerdes.dz

Abstract

Biohydrogen production through dark fermentation represents a promising approach for renewable energy generation and organic waste valorization. However, the complex nonlinear relationships between fermentation conditions and biohydrogen production make accurate prediction challenging. This study develops and evaluates a machine learning framework for predicting cumulative biohydrogen production from the dark fermentation of food waste. A database of 177 experimental observations collected from the literature was established, including fermentation temperature, pH, substrate concentration, fermentation time, and encoded substrate type as input variables. Three supervised learning algorithms, namely Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Regression (SVR), were developed and optimized using Grid Search combined with five-fold cross-validation. Model performance was assessed using the coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), while experimental and predicted kinetic profiles were used for additional visual assessment. SHAP analysis was subsequently applied to the best-performing model to investigate the relative contribution of the main numerical input variables. The results showed that all three models achieved high predictive accuracy, with ANN providing the best performance on the independent test set, achieving an R^2 of 0.9967, an RMSE of 0.3204 mL/L, and an MAE of 0.1641 mL/L. RF also

demonstrated strong predictive capability, whereas SVR exhibited comparatively higher prediction errors. SHAP analysis identified fermentation time as the most influential variable, followed by substrate concentration, pH, and temperature. Overall, the results demonstrate the potential of optimized machine learning models, particularly ANN, for accurate prediction and interpretation of biohydrogen production from food waste dark fermentation.

Keywords: Biohydrogen production; Dark fermentation; Food waste; Machine learning; SHAP.

1. Introduction

Dark fermentative biohydrogen production has emerged as one of the most promising biological routes for sustainable hydrogen generation owing to its ability to convert biodegradable wastes into a clean and renewable energy carrier under relatively mild operating conditions [1,2]. Compared with conventional hydrogen production technologies, dark fermentation offers several advantages, including lower energy requirements, simultaneous organic waste treatment, and reduced greenhouse gas emissions [3,4]. Among the various biodegradable feedstocks investigated for biohydrogen production, food waste has attracted particular attention due to its high biodegradability, abundant availability, and rich carbohydrate content, making it an attractive substrate for large-scale hydrogen production while contributing to circular economy and waste management strategies.

Despite these advantages, the dark fermentation of food waste remains a complex biochemical process influenced by numerous interacting operational and substrate-related parameters [5]. Process variables such as temperature, initial pH, fermentation time, substrate concentration, inoculum characteristics, and the physicochemical composition of food waste collectively affect microbial metabolism and hydrogen production [6]. Moreover, food waste itself is highly heterogeneous, with significant variations in carbohydrate, protein, lipid, and moisture contents depending on its origin and composition [7]. These complex and nonlinear interactions make the prediction of biohydrogen production particularly challenging and limit the applicability of conventional empirical correlations.

Traditionally, biohydrogen production has been investigated using kinetic and

mechanistic models, including first-order kinetics and the Modified Gompertz equation, which have provided valuable insights into microbial growth dynamics and hydrogen evolution [8-11]. Although these approaches remain indispensable for understanding the fermentation mechanism, they generally rely on simplifying assumptions and require parameter estimation procedures that may not adequately capture the complex nonlinear relationships between multiple operational variables [12]. Their predictive capability may therefore decrease when experimental conditions vary significantly across different studies. As a result, data-driven approaches have increasingly attracted attention as effective alternatives for modeling complex fermentation systems.

Recent advances in machine learning (ML) have demonstrated significant potential for predicting biohydrogen production because these techniques can capture complex nonlinear relationships directly from experimental data without requiring explicit mathematical descriptions of the biological process [13, 14]. Several supervised learning algorithms, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forest (RF), have been successfully applied to bioenergy-related applications, providing accurate predictions and fast computational performance [15,16]. Recent years have witnessed a growing interest in applying machine learning techniques to model and predict biohydrogen production from various biological processes. As summarized in Table 1, ANNs have been the most widely adopted approach owing to their ability to capture the nonlinear behavior of fermentation systems. More recently, other supervised learning algorithms, including SVM and RF, have also demonstrated competitive or even superior predictive performance under certain operating conditions. Despite these encouraging advances, existing investigations have generally focused on individual machine learning models or specific fermentation systems, while comprehensive comparative evaluations of optimized machine learning algorithms using a unified and heterogeneous biohydrogen dataset remain limited. This highlights the need for further studies aimed at systematically assessing different predictive models under the same experimental framework.

Despite the growing application of machine learning in biohydrogen prediction, most existing studies have focused on generalized models developed using datasets composed of diverse organic substrates. While these models provide broad applicability, they may

not fully capture the specific characteristics of individual feedstocks. In particular, food waste exhibits distinct physicochemical properties and biodegradation behavior that strongly influence hydrogen production. Consequently, developing machine learning models specifically trained on food waste datasets may improve prediction accuracy while providing reliable tools for process optimization. Furthermore, comparative evaluations of different machine learning algorithms using a unified food waste database remain limited.

Motivated by these challenges, the present study develops a machine learning framework for the accurate prediction of biohydrogen production from the dark fermentation of food waste. A comprehensive database was compiled exclusively from published experimental studies involving food waste under various operating conditions. Three supervised learning algorithms, namely Artificial Neural Networks (ANN), Random Forest (RF), and Support Vector Regressor (SVR), were developed and comparatively evaluated using five-fold cross-validation and multiple statistical performance metrics. In addition, SHAP (SHapley Additive exPlanations) analysis was employed to interpret the predictions of the best-performing model and identify the most influential operational variables affecting biohydrogen production.

Table 1 Representative studies on machine learning applications for biohydrogen production prediction and optimization.

Reference	Biohydrogen process / Feedstock	Machine learning models	Main objective	Main findings
[17]	Photo-fermentation	ANN	Prediction of hydrogen production under different operating conditions.	ANN accurately predicted hydrogen production and successfully generalized from laboratory to outdoor conditions.
[18]	Confectionery wastewater	ANN	Prediction of hydrogen yield	The ANN model achieved excellent predictive performance using COD, pH, and volatile fatty acids as input variables.
[19]	Dark fermentation of coffee mucilage	ANN, Fuzzy Logic	Comparative prediction of biohydrogen production	The fuzzy logic model provided higher prediction accuracy than ANN for the investigated dataset.
[20]	Dark fermentation	ANN, SVM	Modeling and optimization of hydrogen production	SVM produced more accurate predictions than ANN, demonstrating its suitability for nonlinear fermentation systems.
[21]	Dark fermentation	ANN, SVM, RF	Comparative evaluation of machine learning	Random Forest achieved the highest prediction accuracy among the evaluated

				algorithms	models.
[22]	Photocatalytic hydrogen production from wastewater	Multiple models	ML	Performance comparison of predictive models	Bayesian Regularized Multilayer Perceptron exhibited the best predictive performance.
[23]	Dark fermentation	ANN coupled with Genetic Algorithm		Process optimization	Hybrid ANN-GA effectively optimized operating conditions and enhanced hydrogen production compared with conventional optimization methods.
[24]	Organic waste dark fermentation	ANN		Prediction of biohydrogen production from heterogeneous substrates	ANN successfully predicted biohydrogen production from diverse organic wastes, demonstrating good generalization capability.
[25]	Organic waste dark fermentation	BO-ANN-Gompertz		Generalized prediction framework for cumulative biohydrogen production	BO-ANN-Gompertz successfully predicted cumulative biohydrogen production from diverse organic wastes, demonstrating impressing generalization capability.

2. Materials and Methods

The methodology adopted in this study was designed to develop reliable predictive models for estimating cumulative biohydrogen production from the dark fermentation of food waste. The different stages of the study were organized into a sequential workflow, beginning with the collection and preparation of experimental data and followed by model development, hyperparameter optimization, performance evaluation, and model interpretation. Three machine learning approaches, namely Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Regression (SVR), were considered and evaluated under the same methodological framework. A schematic overview of the complete workflow is presented in Figure 1.

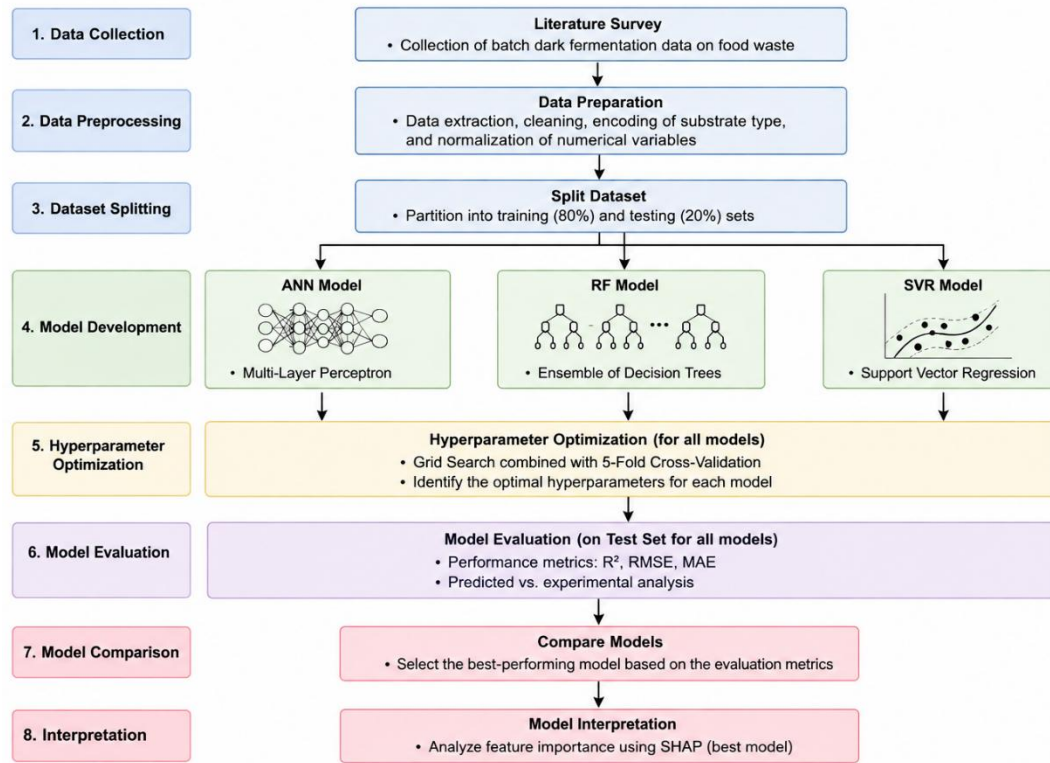


Figure 1 Methodological framework adopted for the development and evaluation of the machine learning models.

2.1. Data collection

The dataset used in this study was compiled from six peer-reviewed articles published between 2015 and 2023 [26-31], reporting batch dark fermentation experiments using food waste as the substrate. A total of 177 experimental observations were extracted from the cumulative biohydrogen production curves reported in these studies. Each observation corresponds to the cumulative hydrogen production measured at a specific fermentation time under defined operating conditions.

To ensure the development of a substrate-specific predictive model, only studies investigating food waste were considered. The selected publications provided sufficient experimental information to reconstruct the cumulative hydrogen production profiles together with the corresponding operating conditions. Figure 2 illustrates the data collection procedure, while the characteristics of the selected studies are summarized in Table 2.



Figure 2 Data collection procedure for constructing the machine learning dataset for biohydrogen prediction.

The following inclusion criteria were applied during the literature survey:

- Batch dark fermentation experiments using food waste substrates.
- Availability of cumulative hydrogen production profiles over the fermentation period.
- Clear reporting of the main operating parameters, including temperature, initial pH, substrate concentration, and hydraulic retention time.
- Sufficient experimental information to reconstruct the dataset for machine learning modelling.

Table 2 Statistical summary of the machine learning dataset.

Variable	Unit	Role	Min.	Max.	Mean
Initial pH	–	Input	4.5	7.9	6.3
Temperature	°C	Input	30	55	38.0
Substrate concentration	g COD L ⁻¹	Input	0.26	147.50	46.73
Hydraulic retention time	h	Input	15	250	38.0
Cumulative H ₂ production	mL L ⁻¹	Output	0	2548.5	420.5

As shown in Table 2, the dataset covers a broad range of operating conditions representative of both mesophilic and thermophilic dark fermentation processes. The wide variation in substrate concentration, pH, fermentation time, and cumulative hydrogen production enhances the diversity of the dataset and improves the ability of the machine learning models to learn the underlying nonlinear relationships.

2.2. Exploratory Data Analysis (EDA)

Prior to model development, an exploratory data analysis (EDA) was performed to assess the quality and characteristics of the collected dataset. This step aimed to examine the distribution of the input and output variables, identify potential outliers or inconsistencies, verify the absence of missing values, and evaluate the overall variability of the experimental data.

The exploratory analysis was carried out using Python visualization libraries. As illustrated in Figure 3, the dataset covers a broad range of operating conditions, including different food waste categories, fermentation temperatures, initial pH values, substrate concentrations, and cumulative biohydrogen production levels. No significant missing values or abnormal distributions were observed. The substantial variability of the experimental conditions and hydrogen production values provides a diverse dataset for model training and contributes to improving the generalization capability of the developed machine learning models.

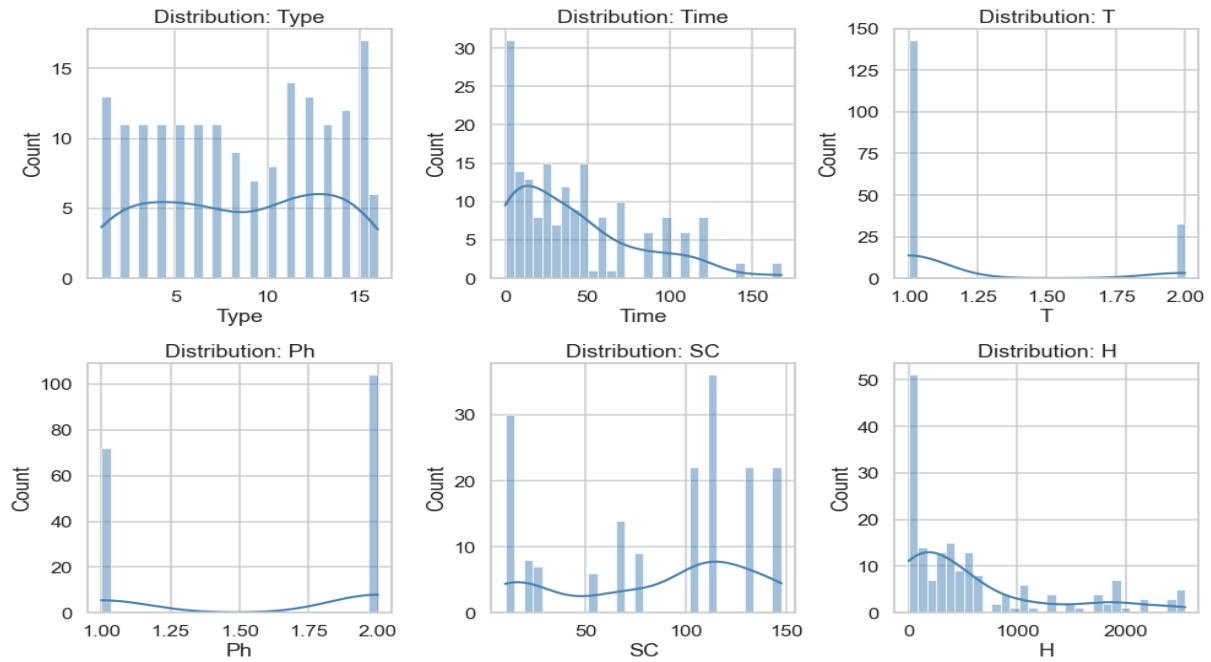


Figure 3 Multidimensional exploratory analysis of the experimental dataset.

2.3. Correlation Analysis

Correlation analysis was performed to examine the linear relationships among the numerical variables included in the dataset prior to machine learning model development. In addition to provide an initial understanding of the data structure, this analysis helps identify potential multicollinearity between predictor variables and offers preliminary insights into their relationships with cumulative biohydrogen production. Pearson's correlation coefficients were calculated for all numerical variables, including fermentation time, temperature, initial pH, substrate concentration, and cumulative hydrogen production.

Figure 4 presents the Pearson correlation matrix as a heatmap, providing a preliminary overview of the linear relationships among the numerical variables. The results indicate that cumulative hydrogen production exhibits a moderate positive correlation with fermentation time, reflecting the progressive accumulation of hydrogen during the fermentation process. Initial pH also shows a positive correlation with hydrogen production, whereas substrate concentration displays a relatively weak correlation, suggesting that its effect may depend on interactions with other operating parameters.

Temperature exhibits a slightly negative correlation within the collected dataset. Overall, no strong multicollinearity was observed among the predictor variables, indicating that each input variable contributes complementary information to the machine learning models.

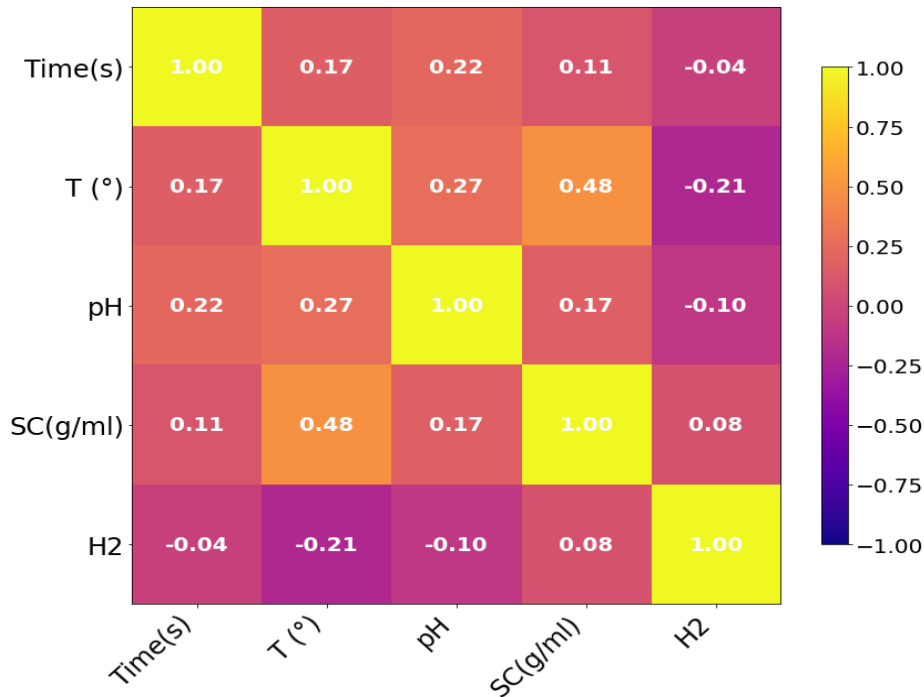


Figure 4 Heatmap of the Pearson correlation matrix for the numerical variables.

Since Pearson correlation only captures linear relationships, the influence of the different variables on model predictions is further investigated using SHAP analysis, which accounts for complex nonlinear interactions. Furthermore, as the food waste type is a categorical variable, it was not included in the Pearson correlation analysis but was incorporated as an input feature during machine learning model development.

2.4. Data preprocessing

Prior to model development, the collected dataset was preprocessed to ensure consistency and compatibility with the machine learning algorithms. Since the data were compiled from multiple literature sources, several preprocessing steps were performed, including the standardization of measurement units, encoding of categorical variables,

and scaling of numerical features. These procedures ensured that all input variables were represented in a consistent format suitable for model training.

2.4.1. Data Standardization

Because the experimental data were collected from different published studies, some variables were reported using different measurement units. To ensure consistency across the dataset, all substrate concentrations were converted to g COD L⁻¹, while cumulative hydrogen production was expressed in mL L⁻¹. Temperature and fermentation time were standardized to °C and hours (h), respectively. The pH values required no conversion because they were already reported on a standardized scale. This standardization step ensured the comparability of experimental observations from different sources.

2.4.2. Categorical Variable Encoding

The food waste type was included as a categorical input variable in the machine learning models. Since machine learning algorithms require numerical inputs, the different food waste categories were transformed into numerical labels using Label Encoding. Each category was assigned a unique integer identifier while preserving its categorical nature. This encoded variable was subsequently used as an input feature during model development together with the numerical operating parameters.

2.4.3. Feature Scaling

To improve model training and ensure numerical consistency among the input variables, the numerical features were normalized using the Min-Max scaling technique. This method transforms each variable into the range [0,1] using Eq. (1).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where x is the original value, x_{min} and x_{max} are the minimum and maximum values of the corresponding feature, and x' is the normalized value.

Feature scaling prevents variables with large numerical ranges from dominating the

learning process and contributes to improve the convergence and predictive performance of machine learning algorithms [.

2.5. Machine Learning Framework

A machine learning framework was developed to predict cumulative biohydrogen production from food waste by dark fermentation using the collected experimental dataset. Three representative supervised learning algorithms, namely Artificial Neural Networks (ANN), Random Forest (RF), and Support Vector Regressor (SVR), were selected because of their ability to model complex nonlinear relationships between operating conditions and hydrogen production.

2.5.1. Model Development and Cross-Validation

The predictive performance of the developed machine learning models was evaluated using a five-fold cross-validation (5-fold CV) strategy. Cross-validation provides a reliable assessment of model generalization by reducing the dependence of the results on a single train-test split while minimizing the risk of overfitting.

In the adopted procedure, the dataset was first divided into a training set (80%) and untouched test set (20%). Then the 5-fold CV was performed only within the training set by randomly dividing it into five approximately equal subsets. During each iteration, four subsets were used for model training, whereas the remaining subset was reserved for validation. This process was repeated five times, allowing each subset to serve once as the validation set. The predictive performance obtained in the five iterations was then averaged to provide a robust estimate of the generalization ability of each machine learning model. A schematic representation of the adopted cross-validation procedure is presented in Figure 5.

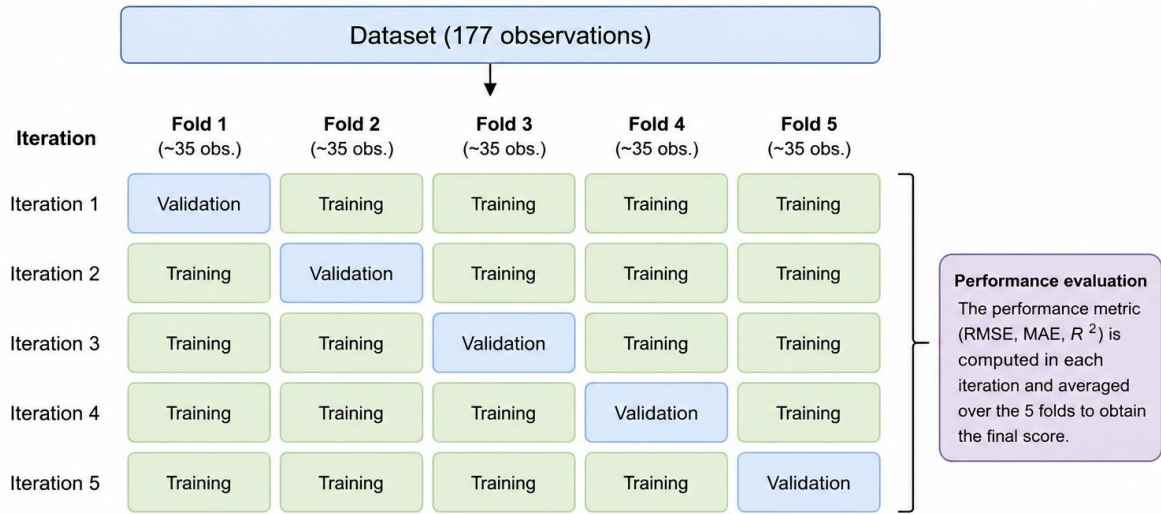


Figure 5 5-Fold Cross validation procedure.

2.5.2. Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are powerful supervised learning algorithms capable of modeling complex nonlinear relationships between input and output variables [32]. An ANN consists of interconnected neurons organized into an input layer, one or more hidden layers, and an output layer. During the training process, the network learns the underlying relationship between the predictor variables and the target output by iteratively adjusting the connection weights to minimize the prediction error. The output of a neuron in a given layer is calculated by applying an activation function to the weighted sum of the outputs from the preceding layer, as expressed by:

$$h_j^{(l)} = f\left(\sum_{i=1}^{nl-1} w_{ij}^{(l)} h_i^{(l-1)} + b_j^{(l)}\right) \quad (2)$$

where $h_j^{(l)}$ is the output of neuron j in layer l , $w_{ij}^{(l)}$ is the weight connecting neuron i of the previous layer to neuron j , $b_j^{(l)}$ is the corresponding bias, and f denotes the activation function.

Owing to their flexibility and strong approximation capability, ANNs have been widely applied to modeling complex biological and biochemical processes, including bioenergy

production [33,34].

In this study, a Multi-Layer Perceptron (MLP) was employed to predict cumulative biohydrogen production from food waste dark fermentation. The model received the encoded food waste type together with the operating parameters as input features, while cumulative hydrogen production was considered as the target variable. The network was trained using the Adam optimization algorithm, and its hyperparameters were optimized through Grid Search combined with the five-fold cross-validation procedure described previously. The optimal ANN configuration was subsequently used for model performance evaluation.

2.5.3. Random Forest (RF)

Random Forest (RF) is an ensemble learning algorithm that combines the predictions of multiple decision trees to improve predictive accuracy and reduce overfitting [35]. Each tree is independently trained using a bootstrap sample of the training dataset, while a random subset of input features is considered at each split. The final prediction is obtained by averaging the outputs of all decision trees using Eq. (3), resulting in a robust model capable of capturing complex nonlinear relationships while maintaining good generalization performance.

$$\hat{y}_{RF}(x) = \frac{1}{N_T} \sum_{K=1}^{N_T} T_k(x) \quad (3)$$

where $T_k(x)$ represents the prediction of the k -th decision tree for input vector x , and N_T is the total number of trees in the forest.

In the present study, the RF model was developed to predict cumulative biohydrogen production from food waste using the encoded food waste type together with the operating parameters as input variables. Bootstrap sampling was employed during training, and the optimal hyperparameters, including the number of trees, maximum tree depth, minimum samples required for node splitting, and minimum samples required at

leaf nodes, were determined using Grid Search combined with the five-fold cross-validation procedure. The optimized RF model was subsequently evaluated and compared with the developed models using the selected performance metrics.

2.5.4. Support Vector Regressor (SVR)

Support Vector Regression (SVR) is a supervised machine learning algorithm derived from Support Vector Machines (SVM) and adapted for regression problems [36]. Unlike conventional regression methods, SVR aims to determine a regression function presented in Eq. (4) that accurately predicts the target variable while maintaining a predefined error tolerance. By employing kernel functions, SVR can effectively capture complex nonlinear relationships between the input variables and the output, making it particularly suitable for modeling nonlinear engineering and biological processes.

$$f(x) = \sum_{i=1}^N (\alpha_i - \alpha_i^*) K(x_i, x) + b \quad (4)$$

where α_i and α_i^* are the Lagrange multipliers, $K(x_i, x)$ is the kernel function, x_i, x represents the support vectors, and b is the bias term.

In this study, the SVR model was developed to predict cumulative biohydrogen production from food waste using the encoded food waste type together with the operating parameters as input variables. The model hyperparameters, including the regularization parameter, kernel type, kernel coefficient, and ε -insensitive loss parameter, were optimized using Grid Search combined with the five-fold cross-validation procedure. The optimized SVR model was subsequently evaluated and compared with the ANN and RF models using the selected performance metrics.

2.5.5. Hyperparameters optimization

The predictive performance of machine learning algorithms strongly depends on the selection of appropriate hyperparameters. Therefore, hyperparameter optimization was performed using the Grid Search approach combined with the five-fold cross-validation

procedure described previously. Grid Search systematically evaluates all possible combinations of predefined hyperparameter values and selects the configuration that provides the best cross-validation performance. The hyperparameter search space considered for the ANN, RF, and SVR models is summarized in Table 3. The ranges presented in Table 03 were discretized into multiple candidate values, and all possible hyperparameter combinations were systematically evaluated using Grid Search with five-fold cross-validation. The optimal hyperparameters identified through Grid Search were subsequently used to train the final ANN, RF, and SVR models.

Table 3 Hyperparameter search space considered for Grid Search optimization of the ANN, RF, and SVR models.

Model	Hyperparameter	Search space
ANN	Hidden layer 1 neurons	5–100
	Hidden layer 2 neurons	5–50
	Activation function (Layer 1)	ReLU, Tanh, Sigmoid
	Activation function (Layer 2)	ReLU, Tanh, Sigmoid
RF	Number of trees ($n_{estimators}$)	50–400
	Maximum tree depth (max_depth)	10–50
	Minimum samples for split ($min_samples_split$)	2–20
	Minimum samples per leaf ($min_samples_leaf$)	1–10
SVR	Regularization parameter (C)	0.01–100
	Epsilon (ϵ)	0.001–0.1
	Kernel coefficient (γ)	0.001–10
	Kernel function	Linear, Polynomial, RBF

The optimal hyperparameter values selected for each model are reported in the Results and Discussion section.

2.6. Model Performance Evaluation

The predictive performance of the developed machine learning models was evaluated using three widely adopted statistical indicators: the coefficient of determination (R^2), the root mean squared error (**RMSE**), and the mean absolute error (**MAE**). These metrics provide complementary information regarding the accuracy and generalization capability of the predictive models.

The coefficient of determination (R^2) measures the proportion of the variance in the

experimental data explained by the model, with values closer to 1 indicating better predictive performance. The RMSE evaluates the average magnitude of the prediction errors while assigning greater weight to large deviations, whereas the MAE represents the average absolute difference between the predicted and experimental values. The three performances metrics are defined as follows:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (7)$$

where y_i and $\hat{y}_i$ represent the experimental and predicted cumulative biohydrogen production, respectively, $\bar{y}$ is the mean of the experimental values, and n is the total number of observations. The values of these metrics were averaged over the five cross-validation folds to provide a robust assessment of model performance.

2.7. Model Interpretation Using SHAP

To improve the interpretability of the predictive models, SHapley Additive exPlanations (SHAP) were employed to analyze the contribution of the input variables to the model predictions. Based on cooperative game theory, SHAP quantifies the individual contribution of each feature to the predicted output, enabling both global and local interpretation of the model.

In this study, SHAP analysis was performed on the best-performing machine learning model identified during the comparative evaluation using Eq. (8). This analysis was used to determine the relative importance of the operating parameters and food waste type in predicting cumulative biohydrogen production and to provide insights into the influence of each input variable on the model predictions.

$$f(x) = E[f(x_0)] + \sum_{m=1}^M \phi_m \quad (8)$$

where $f(x)$ is the final predicted value for a given input x , $E[f(x_0)]$ is the expected model output across the entire dataset (baseline prediction), ϕ_m is the Shapley value corresponding to the m^{th} feature.

3. Results and discussion

This section presents and discusses the results obtained from the developed machine learning models for predicting cumulative biohydrogen production from food waste dark fermentation.

3.1. Hyperparameter optimization

The predictive performance of machine learning algorithms is strongly influenced by the selection of suitable hyperparameters. Therefore, Grid Search combined with five-fold cross-validation was employed to determine the optimal hyperparameter configuration for the ANN, RF, and SVR models. The optimal values obtained after the optimization process are summarized in Table 4.

Table 4 Optimal hyperparameters selected by Grid Search for the developed machine learning models.

Model	Hyperparameter	Optimal value
ANN	Hidden layer 1 neurons	36
	Hidden layer 2 neurons	26
	Activation (Layer 1)	ReLU
	Activation (Layer 2)	Tanh
RF	Number of trees ($n_estimators$)	120
	Maximum tree depth (max_depth)	25
	Minimum samples for split ($min_samples_split$)	6
	Minimum samples per leaf ($min_samples_leaf$)	2
SVR	Regularization parameter (C)	45.3471
	Epsilon (ϵ)	0.02395
	Gamma (γ)	0.0785
	Kernel	RBF

The optimal hyperparameter configurations obtained through the Grid Search procedure are summarized in Table 04. For the ANN model, the optimization process identified an architecture composed of two hidden layers with 36 and 26 neurons employing ReLU and

Tanh activation functions, respectively. This architecture provides sufficient representational capacity to capture the nonlinear relationships between the operating conditions and cumulative biohydrogen production while maintaining a reasonable level of model complexity.

For the RF model, the optimal configuration consisted of 120 decision trees with a maximum depth of 25, a minimum of six samples required to split an internal node, and two samples per leaf. These settings provide a good balance between prediction accuracy and model generalization by limiting excessive tree growth while preserving the diversity of the ensemble.

Regarding the SVR model, the optimization process selected the RBF kernel together with a regularization parameter (C) of 45.3471, an ϵ -insensitive loss parameter of 0.02395, and a kernel coefficient (γ) of 0.0785. The selection of the RBF kernel confirms the nonlinear nature of the relationship between the input variables and biohydrogen production, while the optimized hyperparameters provide an appropriate compromise between model flexibility and generalization capability.

The schematic architectures of the three developed machine learning models are presented in Figures 06, 07, and 08, respectively.

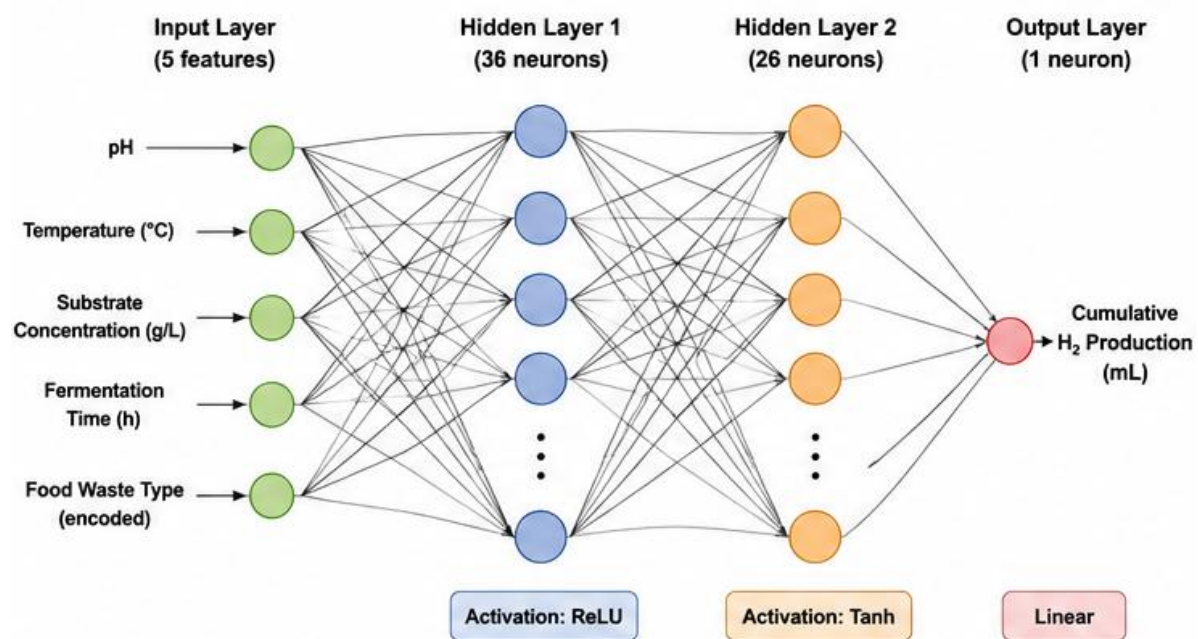


Figure 6 Schematic architecture of the Artificial Neural Network model.

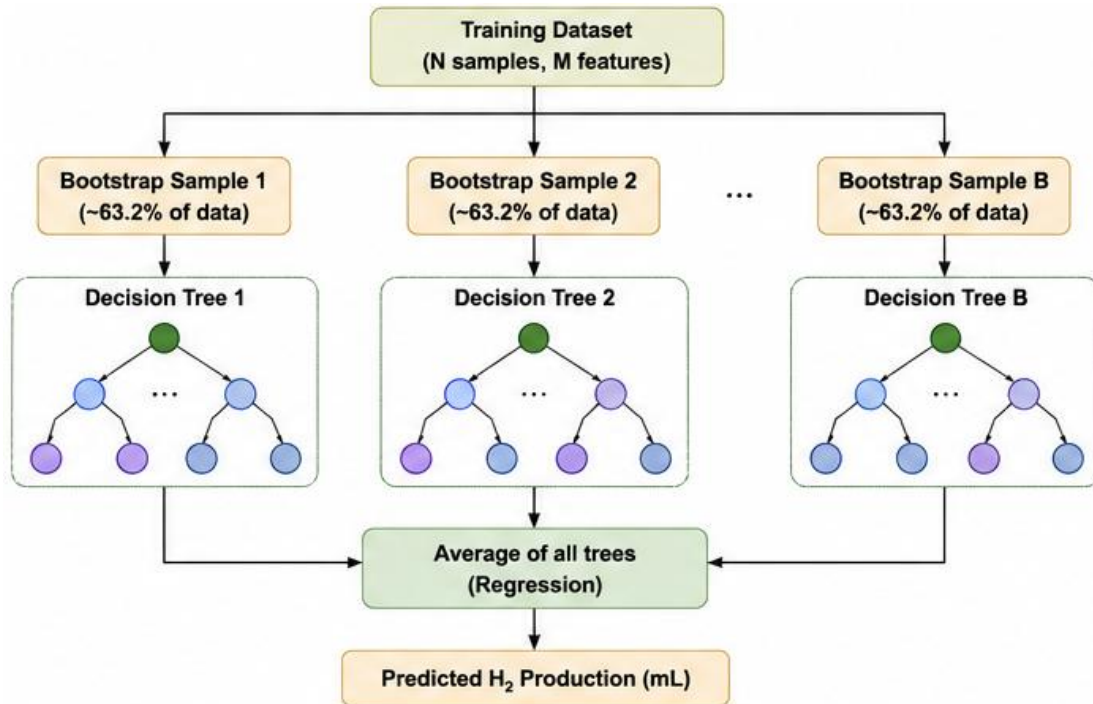


Figure 7 Schematic architecture of the Random Forest model.

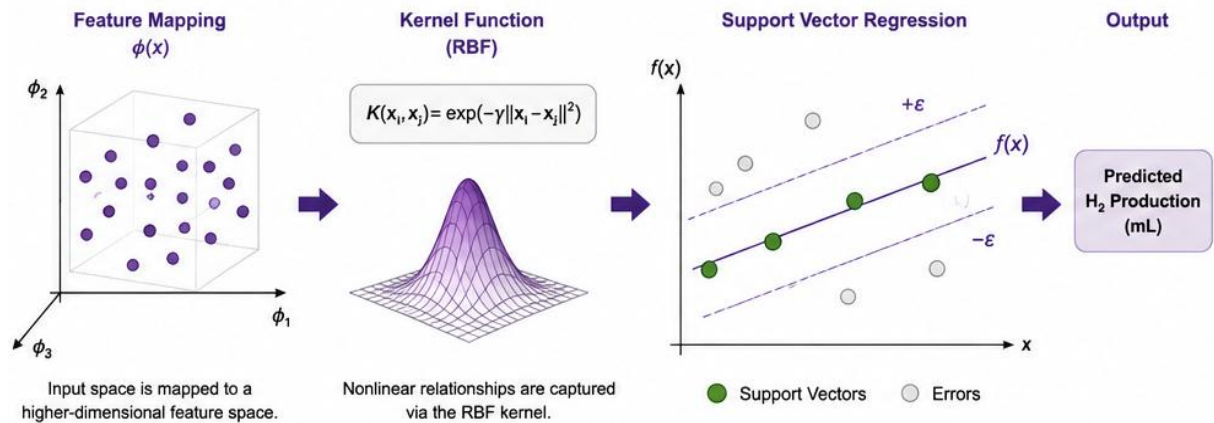


Figure 8 Schematic architecture of the Support Vector Regression model.

3.2. Quantitative assessment of model performance

The predictive performance of the developed ANN, RF, and SVR models was quantitatively assessed using three statistical indicators: the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The quantitative performance of the three developed machine learning models is summarized in Table 05.

Table 5 Performance of the developed ANN, RF, and SVR models for biohydrogen production prediction.

Model	Dataset	MAE (mL/L)	RMSE (mL/L)	R ²
ANN	Training	0.1630	0.3175	0.9985
	Testing	0.1687	0.3319	0.9833
	All data	0.1641	0.3204	0.9967
RF	Training	0.2251	0.4513	0.9570
	Testing	0.2235	0.4376	0.9271
	All data	0.2248	0.4486	0.9580
SVR	Training	0.7132	0.9064	0.9492
	Testing	0.7056	0.8986	0.9146
	All data	0.7116	0.94927	0.9406

Overall, the three models were able to reproduce the experimental cumulative biohydrogen production with different levels of accuracy. However, clear differences in predictive performance were observed among the investigated approaches. The overall performance ranking was ANN > RF > SVR, with the ANN providing the most accurate predictions, achieving the best performance, with an (R²) of 0.9967, an RMSE of 0.3204 mL/L, and an MAE of 0.1641 mL/L. These results indicate a very close agreement between the predicted and experimental cumulative biohydrogen production values.

The RF model ranked second, achieving an (R²) of 0.9580, an RMSE of 0.4486 mL/L, and an MAE of 0.2248 mL/L. Although its performance was lower than that of ANN, the relatively low error values demonstrate that RF was also able to capture the nonlinear relationships between the fermentation conditions and biohydrogen production.

In contrast, SVR provided the lowest predictive accuracy among the three models, with an (R²) of 0.9406, an RMSE of 0.9094 mL/L, and an MAE of 0.7116 mL/L. Nevertheless, the obtained (R²) confirms that SVR was still capable of explaining a substantial proportion of the variability in the experimental data.

The testing results further support the observed performance ranking. ANN maintained a high predictive accuracy on unseen data, achieving an (R²) of 0.9833, an RMSE of 0.3319 mL/L, and an MAE of 0.1687 mL/L. RF and SVR obtained testing (R²) values of 0.9271 and 0.9146, respectively. The relatively small difference between the ANN training and testing errors indicates good generalization, whereas the lower testing (R²) values obtained by

RF and SVR suggest a greater reduction in predictive performance when applied to unseen observations.

Overall, the results demonstrate that the optimized ANN provided the most accurate prediction of cumulative biohydrogen production among the three investigated models, followed by RF and SVR. The superior performance of ANN can be attributed to its ability to represent complex nonlinear interactions among the fermentation parameters through its multilayer architecture.

3.3. *Visual assessment of model performance*

To complement the quantitative evaluation, scatter plots were used to visually compare the predicted and experimental cumulative biohydrogen production values for the ANN, RF, and SVR models. The 45° reference line represents perfect agreement between experimental and predicted values, such that points located closer to this line indicate more accurate predictions.

Figure 9 presents the regression plots comparing the experimental and predicted cumulative biohydrogen production for the three developed models. The dashed 45° line represents the ideal case in which the predicted values exactly match the experimental measurements. Overall, the three models reproduce the general trend of the experimental data, although differences in the degree of dispersion can be observed.

As shown in Figure 9(a), the ANN model exhibits the closest agreement between predicted and experimental values. The training and testing points are strongly concentrated around the 45° line across the entire range of biohydrogen production, with relatively limited dispersion. This visual observation is consistent with the excellent statistical performance of ANN, characterized by the highest R^2 and the lowest RMSE and MAE among the three models.

The RF model in Figure 9(b) also follows the experimental trend reasonably well, particularly at higher production levels. However, a greater dispersion of the points is observed compared with ANN, especially in the lower and intermediate ranges of cumulative biohydrogen production. This greater scatter is consistent with the lower predictive accuracy indicated by its statistical metrics.

As shown in Figure 9(c), the SVR model presents the widest dispersion around the ideal

line. Although the overall relationship between predicted and experimental values remains evident, several observations show noticeable deviations from the 45° line, particularly at low and intermediate production levels. This behavior agrees with the higher RMSE and MAE and the lower R^2 obtained for SVR.

Overall, the regression plots visually confirm the quantitative evaluation and support the performance ranking of ANN > RF > SVR, with ANN providing the closest agreement between predicted and experimental cumulative biohydrogen production.

(a)

(b)

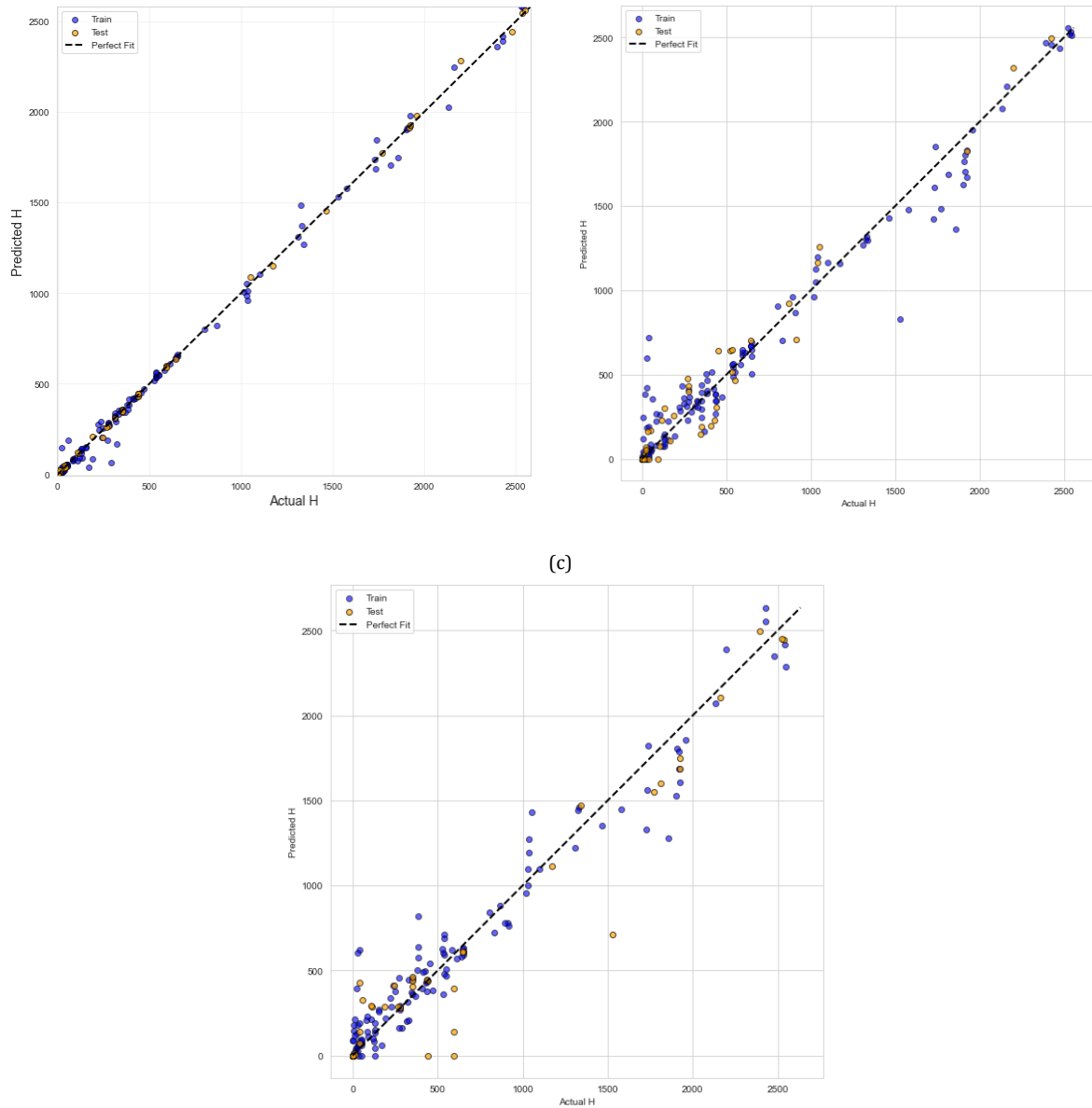


Figure 8 Regression plots of predicted versus experimental cumulative biohydrogen production for (a) ANN, (b) RF, and (c) SVR models.

3.4. Distribution of prediction errors

The distribution of prediction errors was further analyzed to assess the presence of systematic bias and to complement the statistical evaluation of the developed models. The residual error was calculated as the difference between the predicted and experimental cumulative biohydrogen production values:

$$[\text{Error} = H_2 (\text{predicted}) - H_2 (\text{experimental})]$$

(a)

(b)

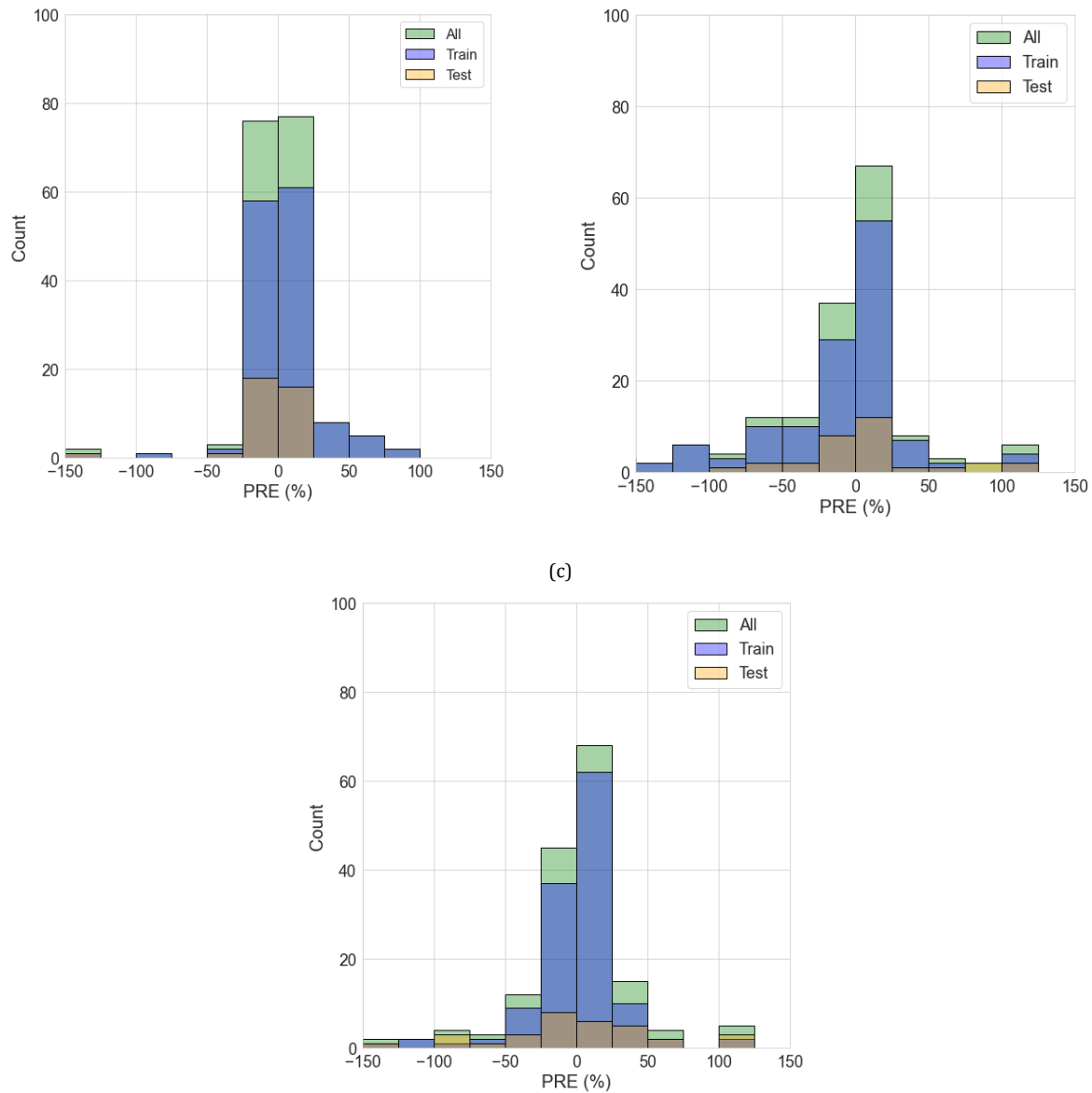


Figure 10 Regression plots of predicted versus experimental cumulative biohydrogen production for (a) ANN, (b) RF, and (c) SVR models.

Figure 10 presents the distributions of the percent relative error (PRE) obtained for the ANN, RF, and SVR models, shown in Figures 10(a), 10(b), and 10(c), respectively.

The three models show their highest concentration of errors around 0%, indicating that no pronounced systematic bias toward overprediction or underprediction is present. However, differences in the spread and tails of the distributions are clearly observed. ANN (Figure 10a) exhibits the most concentrated distribution around zero, with most errors located within a relatively narrow range and only a few extreme observations. This indicates a high degree of prediction consistency and agrees with its lowest MAE and

RMSE values.

The RF model (Figure 10b) also shows a relatively high concentration around zero, but with a broader distribution and more pronounced tails, indicating greater variability in its prediction errors. The SVR model (Figure 10c) presents the widest error distribution, with several observations showing relatively large positive and negative errors. This broader spread is consistent with its higher MAE and RMSE and indicates less consistent prediction accuracy compared with ANN and RF.

Overall, the error distributions complement the quantitative assessment and confirm the performance ranking of ANN > RF > SVR, while also showing that the ANN predictions are the most concentrated and consistent around the experimental values.

3.5. SHAP Analysis for Model Interpretability

Since the ANN model achieved the best predictive performance among the three evaluated algorithms, SHAP (SHapley Additive exPlanations) analysis was applied specifically to this model to provide an interpretable description of the factors driving its predictions. The objective was not to compare the interpretability of the three algorithms, but rather to explain the behavior of the selected best-performing model and identify the relative contribution of the main operating parameters to biohydrogen production. The SHAP analysis was performed using the four numerical input variables: fermentation time (Time), substrate concentration (SC), pH, and temperature (T). The categorical variable representing the encoded substrate type was not included in this interpretability analysis. SHAP values were used to evaluate both the relative importance of the variables and the direction of their contributions to the ANN predictions.

3.5.1. Global Feature Importance

Figure 11 presents the global importance of the input variables based on their mean absolute SHAP values (mean |SHAP|). Fermentation time showed the highest contribution, with a mean |SHAP| value of 400, followed by, substrate concentration (350), pH (210), and temperature (120).

The dominant influence of Fermentation time highlights the importance of the temporal evolution of cumulative hydrogen production. Substrate concentration also shows a substantial contribution, reflecting its role in determining the amount of available organic matter and consequently the potential hydrogen yield. In comparison, pH and

temperature exhibit lower mean absolute SHAP values, suggesting a comparatively smaller contribution to the variation in the ANN predictions within the range of experimental conditions represented in the dataset.

These results highlight the importance of accounting for the heterogeneity of food waste when developing predictive models for dark fermentation.

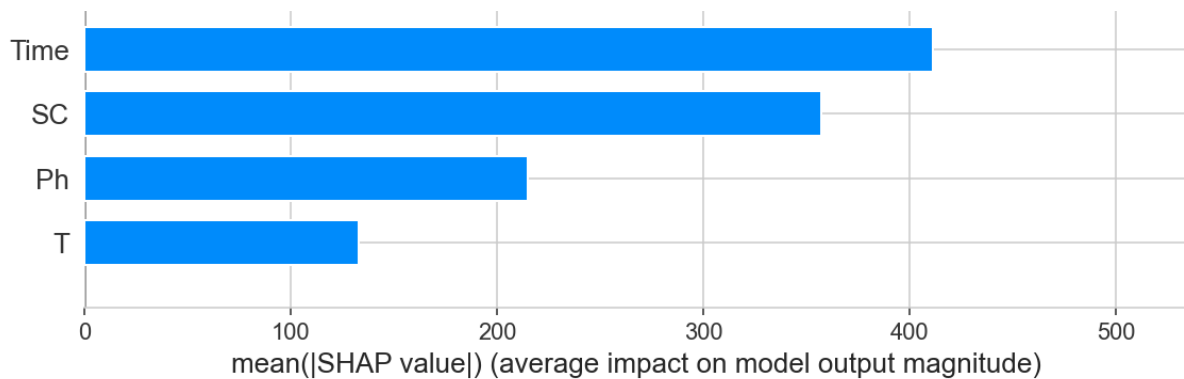


Figure 11 Global feature importance based on mean absolute SHAP values (mean |SHAP|).

3.5.2. SHAP Summary Analysis

The SHAP summary plot presented in Figure 12 provides information on both the magnitude and direction of the contribution of each input variable. Positive SHAP values indicate a contribution toward higher predicted biohydrogen production, whereas negative values indicate a contribution toward lower predictions.

Fermentation time exhibits the widest range of SHAP values and therefore the strongest influence on the model predictions. It shows an overall positive tendency for higher values, with higher time values generally associated with positive SHAP values, indicating that longer fermentation times tend to contribute to higher cumulative hydrogen production. Substrate concentration exhibits a highly dispersed SHAP distribution, with both positive and negative contributions across different concentration levels, suggesting a nonlinear influence. The pH also presents variable contributions depending on its value, with higher pH values generally associated with positive contributions and lower values tending toward negative contributions. In comparison, the SHAP values associated with temperature are more concentrated around zero, indicating a comparatively lower influence on the model predictions.

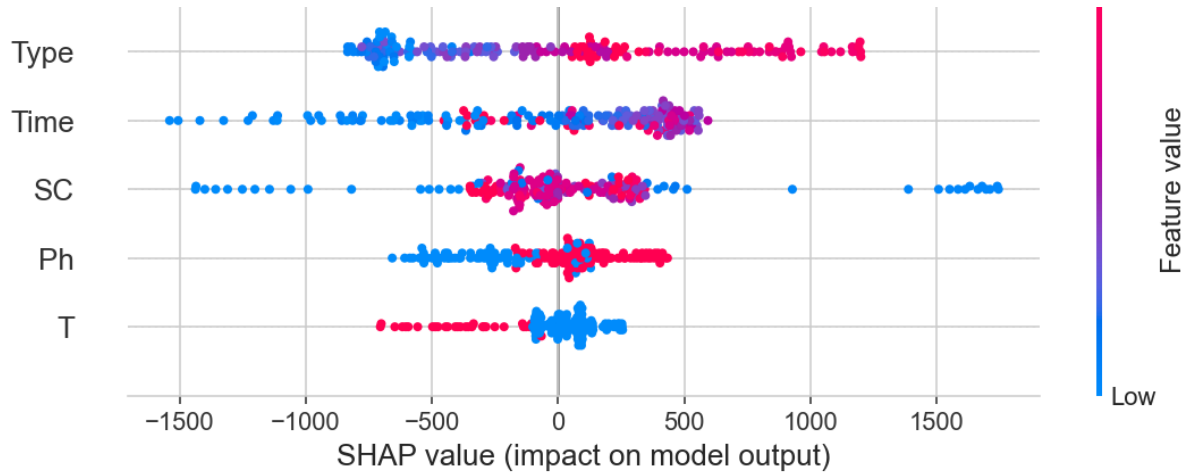


Figure 12 SHAP summary plot showing the contribution of the input variables to ANN predictions.

4. Conclusion

This study developed and evaluated a machine learning-based framework for predicting cumulative biohydrogen production from the dark fermentation of food waste. An experimental database was established from data reported in the literature, integrating different foodwaste types and operating conditions. The collected data were systematically preprocessed through unit standardization, categorical encoding, and numerical normalization to obtain a consistent dataset suitable for machine learning applications.

Three supervised learning models, namely Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Regression (SVR), were developed and optimized using Grid Search combined with five-fold cross-validation. The results demonstrated that all three models were capable of reproducing the experimental biohydrogen production data with high accuracy. However, the ANN model provided the best overall predictive performance, achieving an (R^2) of 0.9833, an RMSE of 0.3319 mL/L, and a MAE of 0.1687 mL/L on the independent test set. RF also showed good predictive capability, while SVR exhibited comparatively higher prediction errors.

The visual evaluation based on the regression plots and error distributions further confirmed the quantitative assessment. The ANN model showed the closest agreement between predicted and experimental values, with a more concentrated distribution of errors compared with RF and SVR. These results further support the suitability of ANN for predicting cumulative biohydrogen production from food waste under the range of

experimental conditions considered in this study.

To provide insight into the factors governing the ANN predictions, SHAP analysis was subsequently performed. Type of substrate and Fermentation time were identified as the most influential variables, followed by substrate concentration, pH, and temperature. The SHAP results also revealed nonlinear effects of several operating parameters, highlighting the ability of the ANN to capture complex relationships that may not be adequately described by conventional correlation analysis.

Overall, the results demonstrate that the proposed machine learning approach can provide accurate and interpretable predictions of cumulative biohydrogen production from dark fermentation of food waste. The developed framework may support the preliminary assessment and optimization of fermentation conditions while reducing the need for extensive experimental trials. It should be noted that the lower performance observed for RF and SVR does not imply an inherent limitation of these algorithms for biohydrogen prediction, as their performance depends on the dataset characteristics, selected hyperparameter search space, and optimization strategy. Alternative optimization approaches and broader search spaces may further improve their predictive performance.

Future work should focus on expanding the database with additional experimental studies and incorporating further biochemical and process-related parameters to improve the generalizability and practical applicability of the predictive framework.

References

- [1] Hovorukha, V., Havryliuk, O., Gladka, G., Tashyrev, O., Kalinichenko, A., Sporek, M., & Dołhańczuk-Śródka, A. (2021). Hydrogen dark fermentation for degradation of solid and liquid food waste. *Energies*, 14(7), 1831.
- [2] Zong, W., Yu, R., Zhang, P., Fan, M., & Zhou, Z. (2009). Efficient hydrogen gas production from cassava and food waste by a two-step process of dark fermentation and photo-fermentation. *Biomass and bioenergy*, 33(10), 1458-1463.
- [3] D'Silva, T. C., Khan, S. A., Kumar, S., Kumar, D., Isha, A., Deb, S., ... & Semple, K. T. (2023). Biohydrogen production through dark fermentation from waste biomass: Current status and future perspectives on biorefinery development. *Fuel*, 350.
- [4] Ntaikou, I., Antonopoulou, G., & Lyberatos, G. (2010). Biohydrogen production from biomass and wastes

via dark fermentation: a review. *Waste and Biomass Valorization*, 1(1), 21-39.

- [5] Ziara RMM, Miller DN, Subbiah J, Dvorak BI (2019) Lactate wastewater dark fermentation: the effect of temperature and initial pH on biohydrogen production and microbial community. *Int J Hydrogen Energy* 44:661–673
- [6] Usman M, Kavitha S, Kannah Y, et al (2021) A critical review on limitations and enhancement strategies associated with biohydrogen production. *Int J Hydrogen Energy* 46:16565–16590
- [7] Kumari, R., Singh, A., Sharma, R., & Malaviya, P. (2024). Conversion of food waste into energy and value-added products: a review. *Environmental Chemistry Letters*, 22(4), 1759-1790.
- [8] Wang J, Wan W (2009) Kinetic models for fermentative hydrogen production: a review. *Int J Hydrogen Energy* 34:3313–3323.
- [9] Gadhe A, Sonawane SS, Varma MN (2014) Kinetic analysis of biohydrogen production from complex dairy wastewater under optimized condition. *Int J Hydrogen Energy* 39:1306–1314.
- [10] Pasupuleti SB, Venkata Mohan S (2015) Acidogenic hydrogen production from wastewater: process analysis with the function of influencing parameters. *Int J Energy Res* 39:1131–1141.
- [11] Karthikeyan B, Gokuladoss V (2022) Fusion of vermicompost and sewage sludge as dark fermentative biocatalyst for biohydrogen production: a kinetic study. *Energies (Basel)* 15:6917.
- [12] Boshagh F, Rostami K, Van Niel EWJ (2022) Application of kinetic models in dark fermentative hydrogen production—A critical review. *Int J Hydrogen Energy* 47:21952–21968.
- [13] Mondal PP, Galodha A, Verma VK, et al (2023) Review on machine learning-based bioprocess optimization, monitoring, and control systems. *Bioresour Technol* 370:128523
- [14] Sultana N, Hossain SMZ, Aljameel SS, et al (2023) Biohydrogen from food waste: modeling and estimation by machine learning based super learner approach. *Int J Hydrogen Energy* 48:18586–18600
- [15] Chong JWR, Khoo KS, Ting H-Y, et al (2025) Artificial Intelligence (AI) approach for the quantification of C-phycocyanin in *Spirulina platensis*: Hybrid stacking-ensemble model based on machine learning and deep learning. *Ecol Inform* 103186
- [16] Ramandani AA, Chong JWR, Srinuanpan S, et al (2025) Artificial intelligence-driven prediction models for the cultivation of *Chlorella vulgaris* FSP-E in food waste culture medium: A comparative analysis and validation of models. *Algal Res* 86:103935
- [17] Monroy I, Guevara-López E, Buitrón G (2018) Biohydrogen production by batch indoor and outdoor photo-fermentation with an immobilized consortium: a process model with Neural Networks. *Biochem Eng J* 135:1–10
- [18] Yogeswari MK, Dharmalingam K, Mullai P (2019) Implementation of artificial neural network model for continuous hydrogen production using confectionery wastewater. *J Environ Manage* 252:109684.
- [19] Moreno Cardenas EL, Zapata-Zapata AD, Kim D (2020) Modeling dark fermentation of coffee mucilage wastes for hydrogen production: Artificial neural network model vs. fuzzy logic model. *Energies (Basel)* 13:1663
- [20] Mahata C, Ray S, Das D (2020) Optimization of dark fermentative hydrogen production from organic

wastes using acidogenic mixed consortia. *Energy Convers Manag* 219:113047.

[21] Zhao S, Li J, Chen C, et al (2021) Interpretable machine learning for predicting and evaluating hydrogen production via supercritical water gasification of biomass. *J Clean Prod* 316:128244.

[22] Kanthasamy R, Ali I, Ayodele BV, Maddah HA (2023) Bio-hydrogen production from the photocatalytic conversion of wastewater: Parametric analysis and data-driven modelling using nonlinear autoregressive with exogeneous input and back-propagated multilayer perceptron neural networks. *Fuel* 344:128026.

[23] Anand A, Mahata C, Moholkar VS (2024) Biohydrogen synthesis from food waste hydrolysate: Optimization using statistical design of experiments (DoE) and artificial neural network (ANN). *Biomass Bioenergy* 191:107452.

[24] Shomope I, Tawalbeh M, Al-Othman A, Almomani F (2025) Predicting biohydrogen production from dark fermentation of organic waste biomass using multilayer perceptron artificial neural network (MLP-ANN). *Comput Chem Eng* 192:108900

[25] Mougari, N. E., Gherzi, D. E., Iachachene, F., Largeau, J. F., & Arici, M. (2026). A generalized hybrid machine learning framework for predicting biohydrogen production via dark fermentation of organic wastes. *Bioprocess and Biosystems Engineering*, 49(2), 351-372.

[26] Turhal, S., Turanbaev, M., & Argun, H. (2019). Hydrogen production from melon and watermelon mixture by dark fermentation. *International Journal of Hydrogen Energy*, 44(34), 18811-18817.

[27] Martínez-Mendoza, L. J., Lebrero, R., Muñoz, R., & García-Depraet, O. (2022). Influence of key operational parameters on biohydrogen production from fruit and vegetable waste via lactate-driven dark fermentation. *Bioresource technology*, 364, 128070.

[28] Casanova-Mina, A. A., Suárez-Vázquez, S. I., Acuña-Askar, K., Alfaro-Barbosa, J. M., & Cruz-López, A. (2024). Continuous dark fermentation by industrial food wastewater: the effect of hydraulic retention time on hydrogen production and microbial variation. *Biomass Conversion and Biorefinery*, 14(19), 23909-23920.

[29] Jais, F. K. A., Hassan, S. H. A., Zamri, M. F. M. A., Zulkornain, M. F., & Shamsuddin, A. H. (2023, June). Biohydrogen production from kitchen organic waste via effective pre-treatment process of dark fermentation. In *IOP Conference Series: Earth and Environmental Science* (Vol. 1205, No. 1, p. 012021). IOP Publishing.

[30] Arslan, C., Sattar, A., Changying, J., Nasir, A., Ali Mari, I., & Zia Bakht, M. (2015). Impact of pH management interval on biohydrogen production from organic fraction of municipal solid wastes by mesophilic thermophilic anaerobic codigestion. *BioMed research international*, 2015(1), 590753.

[31] Basak, B., Fatima, A., Jeon, B. H., Ganguly, A., Chatterjee, P. K., & Dey, A. (2018). Process kinetic studies of biohydrogen production by co-fermentation of fruit-vegetable wastes and cottage cheese whey. *Energy for Sustainable Development*, 47, 39-52.

[32] Krogh A (2008) What are artificial neural networks? *Nat Biotechnol* 26:195–197.

[33] Mougari, N. E., Largeau, J. F., Himrane, N., Hachemi, M., & Tazerout, M. (2021). Application of artificial neural network and kinetic modeling for the prediction of biogas and methane production in anaerobic

digestion of several organic wastes. *International Journal of Green Energy*, 18(15), 1584-1596.

[34] Gheri, D. E., Mougari, N. E., Loubar, K., Amoura, M., & Desideri, U. (2025). A machine learning and Particle Swarm Optimization approach for desiccant wheel modeling and performance prediction. *Applied Thermal Engineering*, 129084.

[35] M.R. Segal, Machine learning benchmarks and random forest regression, (2004).

[36] Awad, M., Khanna, R. (2015). Support Vector Regression. In: Efficient Learning Machines. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4302-5990-9_4