science, engineering and health studies

Research Article

Science, Engineering and Health Studies
https://lio1.tci-thaijo.org/index.php/sehs
ISSN (Online): 2630-0087

Machine learning and experimental design for optimizing nitrogen-rich extract from cassava leaves via liquid hot water extraction

Sirada Subjalearndee[1], Wannee Tinthongkhob[1], Patcharapuek Pattaramanon[2], Nutchapon Chotigkrai[2], Chonlatep Usaku[3], and Nardrapee Karuna[1]*

[1] Department of Biotechnology, Faculty of Engineering and Industrial Technology, Silpakorn University, Nakhon Pathom 73000, Thailand
[2] Department of Chemical Engineering, Faculty of Engineering and Industrial Technology, Silpakorn University, Nakhon Pathom 73000, Thailand
[3] Research Unit on Sustainable Algal Cultivation and Applications, Bio-Circular-Green-economy Technology & Engineering Center (BCGeTEC), Department of Chemical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand

***Corresponding author**:
Nardrapee Karuna
sanchez_n@su.ac.th

***Received**: 25 February 2025*
***Revised**: 21 September 2025*
***Accepted**: 8 October 2025*
***Published**: 16 December 2025*

***Citation**:
Subjalearndee, S., Tinthongkhob, W., Pattaramanon, P., Chotigkrai, N., Usaku, C., & Karuna, N. (2025). Machine learning and experimental design for optimizing nitrogen-rich extract from cassava leaves via liquid hot water extraction. Science, Engineering and Health Studies, 19, 25040011.

ABSTRACT

Cassava leaves are a significant source of nitrogen; however, the severity of the physicochemical extraction processes negatively affects nitrogen release. The objective of this study was to enhance nitrogen-rich extract recovery from cassava leaves through a comparative analysis of various experimental designs and machine learning (ML) techniques. Using the Plackett–Burman design, central composite design, and response surface methodology, the optimal extraction conditions were established: 20 min extraction time, 40% solid loading, and 150 mL extraction volume. The predicted amino nitrogen content reached 209 mg of N, showing a 6% deviation from the experimentally measured value. ML models—specifically, the support vector machine with a radial basis function kernel and random forest (RF)—were subsequently employed to refine the extraction conditions. The RF model showed a 6.6% deviation from the actual value, while both models identified the positive impact of increased solid loading on the total nitrogen recovery. These findings suggest that ML approaches offer promising potential for maximizing the amino nitrogen yield from cassava leaves.

Keywords: nitrogen; Plackett–Burman design; central composite design; support vector machine with radial basis function kernel; random forest

1. INTRODUCTION

The demand for cassava has grown steadily over the years. In 2020, global cassava production reached 303 Mt, with Thailand producing 31.1 Mt (Sowcharoensuk, 2023). After harvesting the desirable cassava root, despite their high protein content (*ca.* 20% db.) (Lammens et al., 2012), a large amount of cassava leaves is often left in the field as agricultural waste. The potential use of cassava leaves as a nitrogen/protein source for fermentation processes has recently been reported (Boundy-Mills et al., 2019; Karuna et al., 2025; Karuna et al., 2023). Compared with the traditional nitrogen source in the culture medium, specifically yeast extract, using cassava leaf extract as the

nitrogen source in *Saccharomyces cerevisiae* culture significantly enhances ethanol production (Karuna et al., 2023). As cassava leaves constitute lignocellulosic biomass, with proteins encapsulated within the structure alongside cellulose, hemicellulose, and lignin (Boundy-Mills et al., 2019; Karuna et al., 2023), an effective extraction method needs to be studied and developed to obtain an extract rich in proteins and nitrogen content.

Although several physicochemical methods, including those with either a diluted acid (Ashokkumar et al., 2022) or an alkaline condition (Karuna et al., 2014; Kim et al., 2016), have been previously used for extracting non-cellulosic materials from lignocellulosic biomass, liquid hot water (LHW) extraction remains the most promising alternative for this specific application (Karuna et al., 2023). This method employs water in its liquid state under high pressure at elevated temperatures (110–230°C) (Aftab et al., 2019; Ruiz et al., 2020), in which it undergoes ionization forming hydroxide ions and hydronium ions (Ruiz et al., 2020; Tomás-Pejó et al., 2011) to penetrate lignocellulosic materials. It triggers the release of acetate from xylan, inducing the hydrolysis and solubilization of hemicellulose, as well as a partial breakdown of lignin (Jönsson & Martín, 2016). Because it does not require chemical addition, LHW extraction produces no toxic wastewater, and subsequent treatment for the remaining toxic chemicals is not required. As these findings reveal, LHW was employed to obtain extracts from cassava leaves in this study.

The optimal conditions of the LHW extraction must be determined to ensure maximum nitrogen content in the cassava leaf extracts obtained, and the extraction is subsequently performed at these conditions. Nonetheless, this is not a trivial task by relying on one-variable-at-a-time experimentation as the extraction has several operating variables with wide working ranges. These include temperature (120–200°C) (Suriyachai et al., 2020), duration (10–60 min) (Olawuni et al., 2024), particle size (50–212 μm) (Kammoun et al., 2023), and solid loading (10–40 mass%) (Zhang et al., 2023). In addition, the pH value is often carefully managed to optimize the efficiency of the process.

Design of experiment (DOE) has been widely used for process optimization across industries (Öğütcü et al., 2024); it systematically explores the combined effects of variables on responses by generating experimental matrices and conducting experiments. Commonly used procedures include the Plackett–Burman design (PBD) for screening influential factors (Cavazzuti, 2013; Plackett & Burman, 1946) and central composite design (CCD) with response surface methodology (RSM) for optimizing variable levels within constraints (Box & Wilson, 1951; Lamidi et al., 2022). Analysis of variance (ANOVA) statistically tests the model's goodness of fit, significance, and prediction errors, identifying significant variable effects.

Machine learning (ML) has recently gained interest in various fields of study because it can handle complex, nonlinear relationships between variables and outcomes—especially when multiple objectives must be simultaneously achieved (Saha et al., 2024). This field includes a range of computational methods that allow systems to learn from data patterns without requiring explicit programming for every situation. The strength of ML lies in its ability to process large amounts of data and identify refined patterns that traditional statistical methods might overlook (Arboretti et al., 2022). This capability is transforming how researchers and engineers approach optimization problems across several disciplines. Furthermore, in situations where the amount of data is enormous or "big data," it outperforms the traditional statistical methods (Vinitha et al., 2023); therefore, ML has rapidly spread across various industries. Recent examples of using ML include the optimization of microalgae culture (Coşgun et al., 2021), in which the existing experimental data were analyzed using the decision tree algorithm to determine the optimal conditions for maximizing biomass growth and lipid yield. ML algorithms, including random forest (RF), support vector machines (SVMs), and gradient boosting, were employed to predict the properties of cellulose-rich materials based on the raw material characteristics and extraction conditions (Phromphithak et al., 2021). ML techniques were used to optimize and predict the bio-crude yield and higher heating values from the hydrothermal liquefaction of lignocellulosic biomass. The critical parameters, including the reaction time, temperature, and pressure, were evaluated. Using the Shapley value method, temperature was identified as the most significant variable. The extreme gradient boosting algorithm provided the most accurate predictions, with deviations of only 5–8% from the experimental results for yields and calorific values (Katongtung et al., 2024). Nonetheless, data are often scarce, especially when technologies/methods of interest have been introduced, making data generation and collection essential for developing ML models. As DOE specializes in generating a necessary set of experiments for exploring the impact of process variables, a joint application of DOE and ML, in which DOE is used as data generation/collection and ML plays a role in the analysis of DOE data, has been recently introduced. Recent studies have shown successful results from the joint applications of DOE and ML. RF with face-centered CCD (FCCD) was used to optimize the cutting-edge ultrasound-aided solvent extraction process. The optimized parameters from ML were well aligned with those from the DOE with RSM (Petchimuthu et al., 2023). The optimal conditions for the manufacture of 316L stainless steel specimens were determined through selective laser melting based on the joint application of RSM with five ML techniques (La Fé-Perdomo et al., 2022).

Although previous studies have described methods for extracting nitrogen from cassava leaves (Karuna et al., 2023), an optimal condition for the LHW extraction of cassava leaves has not yet been reported. Thus, this study aimed to determine the optimal extraction conditions for maximizing the amino nitrogen content in the resulting cassava leaves extract by using DOE with RSM and ML. Initially, using a PBD, three crucial factors were identified from five common variables affecting the nitrogen content in the obtained extract (output of interest): temperature, duration, particle size, extraction volume, and solid loading. Subsequently, a CCD was performed with the identified factors to determine the optimal conditions based on the obtained statistical model. Finally, the SVM and RF algorithms of ML were used with the experimental data obtained from the PBD and CCD to alternatively suggest the optimal conditions for LHW extraction.

2. MATERIALS AND METHODS

2.1 Cassava leaf handling

Cassava leaves of the Rayong 89 strain were collected and dried in the sun for 2–3 days. To obtain dried leaf powder with a specified particle size (within the range of 50 and 212 μm) for subsequent analysis and experiments, the dried leaves were ground using a hammer mill (Retsch, Germany) and then sieved using a vibratory sieve shaker (Retsch Model AS 200, Germany). The obtained particles were then stored in a Ziplock plastic bag.

2.2 Composition analysis

The chemical composition of the cassava leaves was quantified following the standardized laboratory protocols established by the National Renewable Energy Laboratory. The extractable compounds present in the cassava leaves were extracted using a two-step process involving a Soxhlet apparatus. The leaves were first extracted with deionized water until a clear liquid was obtained; this was followed by a second extraction step using ethanol (Sluiter, et al., 2008b). For ash content determination, the cassava leaf powder (size, 50 μm) was placed in a crucible and heated to 575°C for a minimum of 4 h, before weighing the remaining incombustible component (Sluiter et al., 2008a). To determine lignin and carbohydrate contents, the leaf powder was hydrolyzed with hydrochloric acid to liberate lignin and carbohydrate, which were then quantified using UV-Vis spectroscopy and high-performance liquid chromatography, respectively, following that described in (Sluiter et al., 2012). The total nitrogen content of the cassava leaves was determined using the conventional Kjeldahl method as outlined in a previous study (Thiex et al., 2002).

2.3 Extraction procedure

LHW was employed as the extraction method to obtain the cassava leaf extract. Briefly, cassava leaf powder was mixed with deionized water in a 300-mL static Parr reactor at a specified solid loading (% w/w). The reactor was then heated to a target temperature (°C) in a furnace (Chavachote, Thailand). After incubation for a specified duration (min), the reactor was promptly quenched in an ice-cold water bath. The extract was collected by pouring the mixture onto a Whatman #1 filter within a Buchner funnel, and the liquid was separated through vacuum filtration. The collected liquid was stored for subsequent analysis of the total amino nitrogen content (Section 2.5) at 4°C.

2.4 Experimental design and ML

The experimental design was applied to identify the most influential variables to the response in the extraction process. It was also used to generate experimental data for the construction of a mathematical model describing the relationships between the variables and the response (Section 2.4.2) and ML (Section 2.4.3) approaches. Based on the obtained models, the optimal condition(s) for the extraction process were determined.

2.4.1 PBD

To identify the most influential variables from the five variables studied—temperature (X_1), duration (X_2), particle size (X_3), extraction volume (X_4), and solid loading (X_5)—on the total amino nitrogen content, which was treated as the primary response variable in the extraction process, PBD was performed using Design Expert software (version 23.1.4 64-bit, Stat-Ease, Inc., Minneapolis, MN, USA).

As shown in Tables 1 and S1 (Supplementary data), 12 experimental conditions for the extraction process were proposed based on the high (+) and low (−) levels of each variable. As detailed in Section 2.3, the extraction procedure was performed in duplicate using the designed experimental conditions. The three most significant variables were selected based on the resulting *p*-values and confidence levels of each variable.

Table 1. Range of studied variables in the extraction process for the PBD

Factors		Levels	
(codes)	(actual)	–	+
X_1	Temperature (°C)	120	200
X_2	Duration (min)	10	60
X_3	Particle size (μm)	50	212
X_4	Extraction volume (mL)	50	150
X_5	Solid loading (mass%)	10	30

2.4.2 FCCD

Following PBD, in which the three most influential factors in the extraction process were selected, a CCD with RSM was subsequently used to determine the effects of the selected factors: solid loading (X_1, mass%), extraction volume (X_2, mL), and duration (X_3, min), on the total amino nitrogen content in the resulting extract (Y, mg of N). The codes and ranges used in the CCD are presented in Table 2. Here, a three-factor, five-level CCD with α = 1 or FCCD was performed using Design Expert software, suggesting 17 designed runs consisting of six axial points, eight factorial points, and three replicates at the center point (Table 3). To mitigate the potential influence of any unaccounted-for sequential variations or those stemming from external factors, the experimental trials were conducted in duplicate and in a random sequence. The resulting total amino nitrogen content in the extract (Y) from all the runs was then used to determine its relationship with the studied factors (X_1, X_2, and X_3), which is mathematically expressed using the quadratic equation in Equation 1:

$$Y = \beta_0 + \sum_{i=1}^{3} \beta_i X_i + \sum_{i=1}^{3} \beta_{ii} X_i^2 + \sum_{i=1}^{2} \sum_{j=i+1}^{3} \beta_{ij} X_i X_j \quad (1)$$

where β_0 is a model constant, β_i, β_{ii}, and β_{ij} are the regression coefficients accounting for the linear, quadratic, and interaction effects of the variables on the response Y, respectively. X_i and X_j were estimated using Equation 2, which represents the dimensionless coded values of the i^{th} and j^{th} variables, respectively:

$$X_i = \frac{x_i - x_{i,0}}{\Delta x_i} \quad (2)$$

where x_i is the actual value, $x_{i,0}$ is the actual value at the center of the studied range, and Δx_i is the step change value (Yahya et al., 2023).

To evaluate whether the obtained model adequately describes the experimental data, the statistical parameters—lack of fit, coefficient of determination (R^2),

adjusted R^2, predicted R^2, and adequate precision—were estimated by ANOVA. A two-tailed Student's t-test was used to determine the statistical significance of the variable effects on the response. To visualize the variable effects, surface plots were constructed based on the obtained model. Statistical analysis and surface plot construction were performed using Design Expert software.

Table 2. Range of selected influential factors for the FCCD

Factors		Levels		
(codes)	(actual)	−1	0	1
X_1	Solid loading (mass%)	10	25	40
X_2	Extraction volume (mL)	50	100	150
X_3	Duration (min)	20	30	40

Table 3. Designed runs and results of the FCCD

Run	Space type	X_1	X_2	X_3	Total amino nitrogen content, Y (mg of N) †
1	Axial	−1	0	0	41 (7)
2	Factorial	−1	−1	+1	13 (1.7)
3	Factorial	−1	+1	−1	50 (1.4)
4	Factorial	−1	−1	−1	17 (2.7)
5	Factorial	−1	+1	+1	54 (2.8)
6	Axial	0	+1	0	111 (0.7)
7	Axial	0	−1	0	67 (4.8)
8	Axial	0	0	−1	116 (9.8)
9	Center	0	0	0	92 (2.1)
10	Center	0	0	0	126 (4.7)
11	Center	0	0	0	129 (6)
12	Axial	0	0	+1	102 (2.5)
13	Axial	+1	0	0	187 (5.7)
14	Factorial	+1	+1	−1	221 (9.9)
15	Factorial	+1	−1	−1	79 (3.1)
16	Factorial	+1	+1	+1	175 (4.2)
17	Factorial	+1	−1	+1	100 (7.3)
18	Predicted	+1	+1	−1	209
19	Experiment	+1	+1	−1	221

Note: †Standard deviation of duplicate measurements is indicated in parentheses.

To determine the optimal levels of the factors in which the highest total amino nitrogen content in the extract (Y) would be obtained, the desirability analysis in the Design Expert software was used. Here, the objective function of maximizing the total amino nitrogen content in the extract was applied, and based on the obtained mathematical model, an optimal level(s) of the variables was suggested.

2.4.3 ML

A dataset of 29 experimental data points was collected from the PBD (12 data points) and CCD (17 data points) experiments. Five extraction parameters, including temperature, time, extraction volume, particle size, and solid loading, were set as features. The target was the total extracted nitrogen. Table 4 summarizes the range and unit of each parameter. The features' values were normalized using a standard scaler.

Table 4. Description of the dataset

Parameter	Range	Unit	Step size†
Feature			
Temperature	120–200	°C	10
Size	50–212	µm	6
Duration	10–60	minute	1
Extraction volume	50–150	mL	5
Solid loading	10–40	mass%	1
Target			
Total nitrogen	9–221	mg of N	

Note: †Optimal conditions were searched using these step sizes.

ML models were developed in the Python programming platform (Visual Studio Code) using the Scikit-learn library. A SVM with three kernel functions—linear, polynomial, and radial basis function (RBF)—along with a RF was used to predict the total extracted amino nitrogen. Table 5 provides an overview of the different ML techniques employed in this research. The leave-one-out cross-validation technique was used in this study because of the small dataset (Díez Valbuena et al., 2024). During development, the model was trained with *N-1* samples to predict the remaining one sample as the test sample. The average of 29 times for R^2 and root mean square error (RMSE) were used to evaluate the predictive performance of the model and calculated using Equation 3 and Equation 4, respectively. To minimize the mean RMSE, the hyperparameters of the SVM and RF models were optimized using the grid search technique. The best hyperparameters of the SVM for the linear, polynomial, and RBF were (C = 2,601; epsilon = 0.1; gamma =1), (second degree; C = 277; epsilon = 21; gamma = 0.00038), and (C = 194; epsilon = 0.001; gamma = 0.0009), respectively, while the best hyperparameters of the SVM for the RF were (max_depth = 3 and n_estimators = 9). The models with optimized hyperparameters were trained with 29 data points to predict total nitrogen. The mean absolute percentage error (MAPE) was calculated using Equation 5 for comparison with the existing literature. Shapley additive explanation (SHAP value) was used to show the impact of features on the model output using the SHAP library. The highest total nitrogen was predicted by the selected trained models using the feature interval (Table 4).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Predicted_i - Actual_i)^2}{\sum_{i=1}^{n}(Actual_i - \overline{Actual})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Predicted_i - Actual_i)^2}{n}} \quad (4)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|Predicted_i - Actual_i|}{Actual_i} \quad (5)$$

Table 5. ML algorithms, description, and applications

ML algorithms	Description	Applications
VM	SVM is a powerful technique that is widely used in engineering for classification and regression tasks. This algorithm works by leveraging a training dataset to pinpoint a hyperplane that optimally distinguishes data points into distinct categories, maximizing the margin, which is the greatest distance between the hyperplane and any data point from different classes. This hyperplane is defined by support vectors, which are the data points on the boundaries of the margin. This characteristic is the origin of the name "support vector machine" (Xia, 2020). For regression, the goal is to fit as many input data points as possible within the margin while minimizing margin violations, rather than maximizing the margin between two classes. Initially, SVM is outlined for the linearly separable case as described by the following equation in terms of the support vectors: $y_i(w^T x_i - b) \geq 1$ where y_i is the class value of train data x_i. The vector w represents a normal vector. The vectors x_i are the support vectors. Parameter b determines the hyperplane position (Sittichoksataporn & Choksuriwong, 2012). When linear functions are inadequate for separating two classes, kernel functions are used. Their purpose is to map nonlinear data from a lower-dimensional space to a higher-dimensional space. This transformation is achieved using the following equation: $w = b + \sum \alpha_i y_i K(x_i, x_j)$ where α_i is the Lagrange multiplier. Polynomial and RBF are widely used as kernel functions for SVM (Malek et al., 2019). Polynomial $K(x_i, x_j) = (x_i \cdot x_j)^d$ RBF $K(x_i, x_j) = exp\left(-\gamma\|x_i - x_j\|^2\right)$	Battery state estimation (Manoharan et al., 2022), machine condition monitoring and fault diagnosis (Widodo & Yang, 2007), microscopy (Wang & Fernandez-Gonzalez, 2017), precise agriculture (Kok et al., 2021), solar and wind energy resources forecast (Zendehboudi et al., 2018), and structural reliability analysis (Roy & Chakraborty, 2023)
RF	RF is a rule-based algorithm that integrates multiple decision trees into a forest (with random subsets of features) using the ensemble method. Each tree is trained on a random subset of features, and their predictions are averaged to enhance accuracy and prevent overfitting (Wang & Gao, 2022). It introduces "randomness" into the prediction process by applying bootstrap sampling techniques iteratively.	Prediction model of the pore structure in global shale sediments (Jiang et al., 2023), and prediction of the yield strength of as-cast alloys (Zhang et al., 2024)

3. RESULTS AND DISCUSSION

3.1 Composition of cassava leaves

Cassava leaves of the Rayong 89 variety contained 9.9% w/w ash, 28.9% w/w water extract, and 12.9% w/w ethanol extract. Water and ethanol extracts primarily targeted nonstructural compounds. Water extracts proteins (Karuna et al., 2023), soluble sugars (Zhang et al., 2007), certain phenolic compounds (Mota et al., 2008), saponins, and glycosides (Sparg et al., 2004), as well as water-soluble vitamins—such as vitamin C and some B vitamins—and minerals (Manach et al., 2004). In contrast, ethanol primarily extracts flavonoids, polyphenols, and saponins (Chahyadi & Elfahmi, 2020). Therefore, the extraction results showed that cassava leaves contained higher combined amounts of proteins, soluble sugars, certain phenolic compounds, saponins, glycosides, and water-soluble vitamins than the total amount of flavonoids and polyphenols. The composition also included 13.6% w/w glucan and 7.7% w/w xylan, along with 22.4% w/w acid-insoluble residue. Furthermore, the compost contained 2.94% nitrogen, equivalent to 18% protein, using a conversion factor of 6.25 (Table 6). Despite the influence of factors such as cultivar, leaf position, and plant age on the quality of cassava leaves (Chaiareekitwat et al., 2022), the protein content in this particular strain remained comparable with that in previous studies, ranging from 17% to 34% on dried basic (Chaiareekitwat et al., 2022; Hue et al., 2012).

3.2 Plackett–Burman experimental design

A PBD was first performed to determine the most influential factors on the nitrogen content in the casava leaf extract (Y) among the five studied operating variables of the LHW pretreatment process: temperature (X_1), duration time (X_2), particle size (X_3), extraction volume (X_4), and solid loading (X_5). As shown in Table S1 (see Supplementary Material), the experimental data of the total nitrogen content in the extract were obtained by performing the extraction using the designed experimental

conditions. Statistical linear relationships between the studied factors and the response (Y) were computationally determined through regression and ANOVA. Table 7 shows the obtained key statistical parameters, p-value and confidence level, for each factor, which suggested the extent to which each factor affected the response. The lower the p-value (or the higher the confident level), the more influential the response. X_5 (solid loading), X_4 (extraction volume), and X_2 (duration) were found to be the most important (ranking 1, 2, and 3, respectively), as indicated by their highest confidence levels (or lowest p-values) of 99.99% (p<0.0001), 93.11% (p<0.0689), and 89.2% (p<0.1079), respectively. Although the p-values indicated that the extraction volume and duration were not significant during the screening test, these factors may still prove significant under the optimized conditions explored in the CCD experiments (SixSigma, 2024). Therefore, they were still selected for the subsequent CCD for further determination of their optimal levels toward maximizing the total nitrogen content in the extract.

Table 6. Proximate composition of cassava leaves of the Rayong 89 variety

Composition	Unprocessed leaf (% w/w)†
Moisture	11.1
Ash	9.9 (0.01)
Total extractives	41.8
Extractive in water	28.9
Extractive in ethanol	12.9
Acid-insoluble residue	22.4 (3.86)
Nitrogen	2.9 (0.12)
Structural sugars	22.0
Glucan	13.6 (0.58)
Xylan	7.7 (0.60)
Arabinan	0.7 (0.05)
Total	99.0 (3.96)

Note: †Standard deviation of triplicate measurements is indicated in parentheses.

Table 7. ANOVA table for the response model from the nitrogen concentration

Source		Sum of squares	p-value	Confidence level (%)	Influential ranking
X_1	Temperature (°C)	0.2071	79.29	79.29	5
X_2	Duration (min)	0.1079	89.21	89.21	3
X_3	Particle size (μm)	0.1176	88.24	88.24	4
X_4	Extraction volume (mL)	0.0689	93.11	93.11	2
X_5	Solid loading (mass%)	<0.0001	99.99	99.99	1

3.3 Face-centered composite experimental design

A FCCD with RSM was conducted with the selected most influential factors—solid loading (X_1), extraction volume (X_2), and duration (X_3)—to suggest their optimal level(s) for achieving the highest total nitrogen content in the extract (response, Y). Statistical analysis results and parameters of the model—model significance, lack of fit, R^2, adjusted R^2, predicted R^2, and adequate precision—are presented in Table 8. The plots of residuals versus normal probability, model predictions, and run numbers from the model were evaluated to ensure that the ANOVA result was valid and that they met the assumptions required for ANOVA (Figure S1 in Supplementary data). The obtained model was significant (p = 0.0002) with an insignificant lack of fit (p = 0.6935). Furthermore, the model's R^2 and adjusted R^2 were close to unity (0.9679 and 0.9266, respectively), indicating that the model adequately describes the experimental data. Based on the obtained p-values for each factor, the solid loading (X_1) and extraction volume (X_2) were significant to the total nitrogen content in the extract (p<0.0001 and p = 0.0002, respectively) while the duration (X_3) was insignificant. This was in fair agreement with the PBD result, in which solid loading was found to be the most influential factor, followed by extraction volume, while duration was not previously found significant. The interactions between the factors were not significant; however, the power term of the extraction volume was identified as a significant factor to be included in the coded Equation 6.

$$Y = 113.85 + 68.50X_1 + 52.50X_2 - 26.17X_2^2 \quad (6)$$

An equation derived from the experimental data was represented in the response surface plots, which show the total amino nitrogen extracted from cassava leaves (Figure 1) as a function of two factors simultaneously, while maintaining all other factors constant. The code equation is valuable for assessing the relative influence of the variables by examining the coefficients of each factor. Consequently, based on Equation 6, the solid loadings (X_1) exert a greater influence on the total extract nitrogen than the extraction volume (X_2), and a higher extraction volume has a negative impact on the response.

The optimum condition for amino nitrogen extraction from this experimental design was achieved with a duration of 20 min with 40% solid loading, and the extraction volume was set at 150 mL. The particle size and temperature were kept constant at 212 μm and 120°C, respectively. Under these conditions, the predicted amino nitrogen concentration was 209 mg of N, whereas the actual measured content was 221 mg of N; this represented a 6% deviation from the actual value. The extraction process duration was a key factor influencing the release of nitrogen components from cassava leaves.

Table 8. ANOVA results of the proposed mathematical model based on the FCCD

Source	Sum of squares	Mean square	F-value	*p*-value	Significance*
Model	150.66	16.74	23.43	0.0002	Significant
X_1-Solid loading	102.88	102.88	144.03	<0.0001	Significant
X_2-Extraction volume	33.40	33.40	46.76	0.0002	Significant
X_3-Duration	0.2031	0.2031	0.2843	0.6104	Insignificant
X_1X_2	0.7982	0.7982	1.12	0.3256	Insignificant
X_1X_3	0.0128	0.0128	0.0179	0.8972	Insignificant
X_2X_3	0.4621	0.4621	0.6469	0.4477	Insignificant
X_1^2	0.8879	0.8879	1.24	0.3017	Insignificant
X_2^2	4.34	4.34	6.08	0.0431	Significant
X_3^2	0.1098	0.1098	0.1537	0.7067	Insignificant
Lack of fit	3.12	0.6232	0.6614	0.6935	Insignificant
Statistic parameters					
R^2	0.9679				
Adjusted R^2	0.9266				
Predicted R^2	0.6605				
Adeq. Precision	16.3999				

Note: *Data were considered significant when the p<0.05.

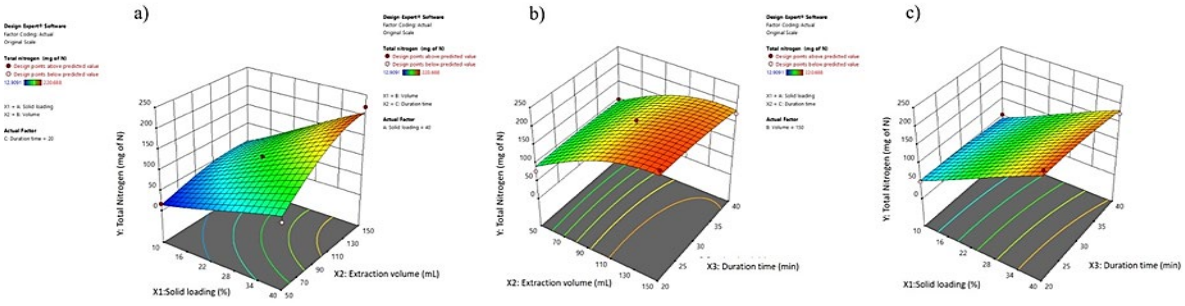

Figure 1. The response surface methodology (RSM) of the face-centered central composite design (FCCD) for optimizing the amino nitrogen concentration
Note: The effects of the independent variables X_1: solid loading, X_2: extraction volume, and X_3: duration on the dependent variable Y: total nitrogen.

The duration allocated for extraction directly impacts the kinetics of nitrogen extraction. An optimal duration ensured that the process reached equilibrium without degrading sensitive components. In this study, the significant influence of the extraction duration highlights the importance of striking a balance to maximize the nitrogen yield while minimizing the potential degradation. The quantity of cassava leaf material used, represented by the solid loadings, was a critical factor in determining the concentration of nitrogen components in the extracted solution. Higher solid loadings contributed to increased nitrogen content. The extraction volume level applied during the extraction process also emerged as a significant factor in this study. Higher extraction volumes increase the solute diffusion driving force from the solid into the solvent. This process is governed by Fick's law of diffusion, where a larger solvent volume reduces the solute concentration in the solvent phase, thereby maintaining a concentration gradient for continuous extraction. For instance, the ethanol ratio and temperature were key variables in the polyphenol extraction from grape pomace experiment. The solvent volume indirectly affects the extraction kinetics by altering the solvent penetration into the solid matrix (Moldovan et al., 2019); however, excessively high volumes can dilute the extract, reduce process efficiency, and increase costs (Azwanida, 2015).

Solid loading and extraction volume were significant factors influencing nitrogen extraction from cassava leaves (Table 8); this is likely because the total nitrogenous compounds available for extraction are determined by solid loading. Higher loading increases substrate availability but requires sufficient solvent for dissolution, and the mass transfer efficiency is governed by the extraction volume. Low volumes limit solute diffusion, reducing the nitrogen yield. Excessive volumes dilute extracts and increase costs without yield improvement (Chahyadi & Elfahmi, 2020).

3.4 Model evaluation and feature impact

The variance between the actual and predicted values was evaluated using the predicted values (Figure 2). This comparison revealed low variance values, with an adjusted R-squared (adjusted R^2) of the model of 0.9266 with a p-value of 0.0002 from the statistical regression method. This outcome suggested that the quadratic model term for cassava leaf extraction was significant. However, the mean R^2 values of the training set for SVM with linear, polynomial, RBF, and RF were within the acceptable ranges of 0.82, 0.82, 0.96, and 0.96, respectively. The mean RMSE of the test set served as a comprehensive metric to evaluate the robustness of the models. The analysis showed that SVM with polynomial regression exhibited the lowest accuracy, with an RMSE of 46.89, whereas the RMSE values for linear, RBF, and RF were 22.89, 17.10, and 13.37, respectively. This disparity implies that the dataset has complex characteristics. MAPE values for SVM with linear, polynomial, RBF, and RF were 29.1%, 13.1%, 13.4%, and 12.2%, respectively. Although the SVM-RBF exhibited excellent performance (the highest R^2 value for the training set and the lowest MAPE), it also showed a high mean RMSE for the test set, indicating the model's fair robustness for unseen data (Montesinos López et al., 2022). Only the SVM-RBF and RF models showed good reliability, with MAPE values between 1% and 16% for ML predictions of solvent extraction (Iweka et al., 2023; Patil et al., 2023; Petchimuthu et al., 2023). This confirms the accuracy of these trained models.

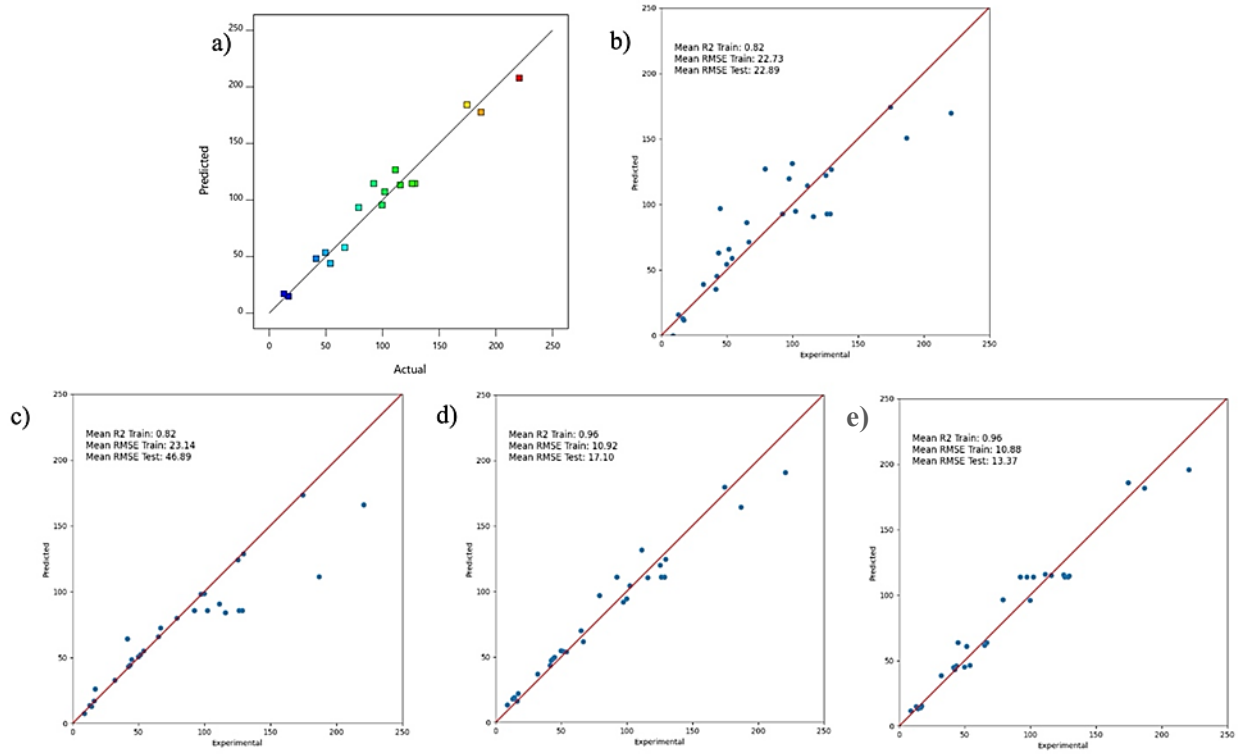

Figure 2. Comparison between predicted and experimental total nitrogen levels
Note: Predictions stem from various models, including a) statistic regression; and SVM models: b) linear regression, c) polynomial regression, d) SVM-RBF, and e) RF.

Table 9 shows the optimal conditions for extracting nitrogen from cassava leaves using trained SVM-RBF and RF models. The highest total nitrogen yields were predicted as 192 and 196 mg of N from the SVM-RBF and RF techniques, respectively. These ML model results closely matched the RSM prediction result of 209 mg of N. The optimum conditions could be more than one, especially in complex systems. After the optimized conditions were validated, the nitrogen contents were found to be 8.9%, 6.6%, and 6% using the SVM-RBF, RF, and statistical methods, respectively. This suggests that RF and statistical regression demonstrated better prediction accuracy than SVM-RBF. One potential reason for this difference in accuracy might be the disparity in the amount of data used for training the ML models. While the SVM-RBF and RF modes were trained using the full dataset of 29 experimental data points, the statistical regression method was applied to a smaller subset of 17 data points for the RSM calculations. Despite having access to the same amount of training data, the inherent differences in the SVM-RBF and RF algorithms may have contributed to their varying levels of prediction accuracy for the amino nitrogen content. The specific characteristics of the dataset and the target variable may have favored the RF approach in terms of predictive performance compared with the SVM-RBF method.

Table 9. Comparison of the optimum conditions obtained from various analyses

Method	SVM-RBF	RF	Statistical regression
Solid loadings (mass%)	40	40	40
Extraction volume (mL)	150	150	150
Duration (min)	24	24	20
Particle size (μm)	212	212	212
Temperature (°C)	120	120	120
Predict-N (mg)	192	196	209
Actual-N (mg)	209	209	221
Off (%)	8.9	6.6	6

The SHAP value plots of the SVM-RBF and RF models showed that the most influential factors for nitrogen extraction were solid loading and extraction volume (Figure 3). Higher solid loading and extraction volume had a positive effect on total nitrogen. These trends were consistent with the statistical regression findings.

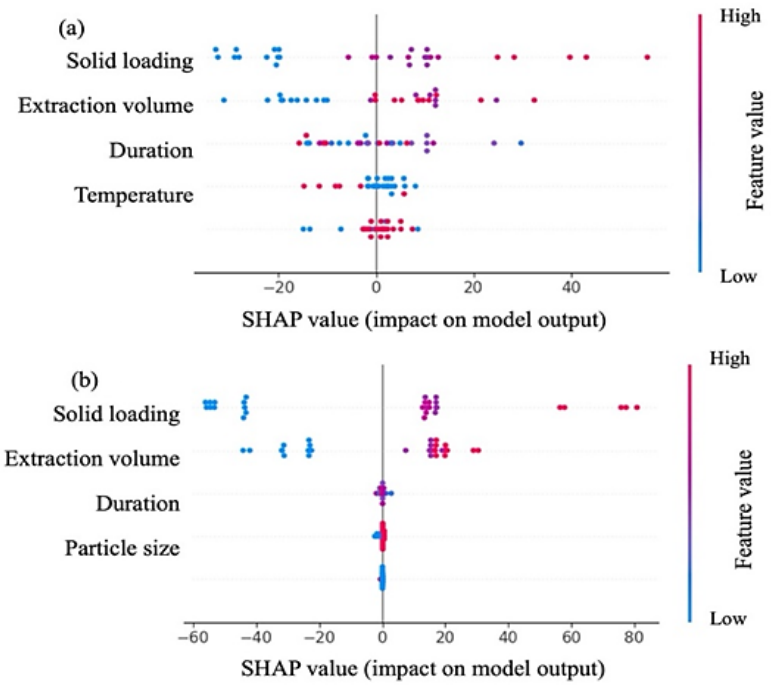

Figure 3. SHAP value plots of (a) SVM-RBF, and (b) RF

4. CONCLUSION

This study successfully determined the optimal conditions for extracting nitrogen from cassava leaves, recognizing its importance as a nitrogen source for fermentation processes. Significant factors influencing amino nitrogen extraction were identified by employing experimental design methodologies such as PBD and FCCD. Through RSM analysis, the optimal extraction conditions were determined, including a 20-min extraction duration, 40% solid loading, and a 150 mL extraction volume. Despite a slight difference between the predicted and experimental values, ML models, specifically RF, effectively predicted the nitrogen content. Furthermore, the study revealed the positive effect of increased solid loading and extraction volume on the total extracted nitrogen. The comparative analysis showed that statistical regression and ML techniques, such as RF, provided accurate predictions of the optimized nitrogen content. The obtained high-quality nitrogen extract can be further explored for its potential applications in various fermentation-based industries, such as biofuel and bioplastic production.

ACKNOWLEDGMENTS

This research was financially supported by the Thailand Science Research and Innovation National Science, Research, and Innovation Fund for fiscal year 2024. The authors would like to express their gratitude to Assistant Professor Kanokwan Teingtham from the Department of Agronomy at Kasetsart University, Thailand, and Mr. Phoominan Jindapang for generously providing the cassava leaves.

REFERENCES

Aftab, M. N., Iqbal, I., Riaz, F., Karadag, A., & Tabatabaei, M. (2019). Different pretreatment methods of lignocellulosic biomass for use in biofuel production. In A. Abomohra (Ed.), *Biomass bioenergy - Recent trends and future challenges* (pp. 15–38). IntechOpen. https://doi.org/10.5772/intechopen.84995

Arboretti, R., Ceccato, R., Pegoraro, L., & Salmaso, L. (2022). Design of experiments and machine learning for product innovation: A systematic literature review. *Quality and Reliability Engineering International*, *38*(2), 1131–1156. https://doi.org/10.1002/qre.3025

Ashokkumar, V., Venkatkarthick, R., Jayashree, S., Chuetor, S., Dharmaraj, S., Kumar, G., Chen, W.-H., & Ngamcharussrivichai, C. (2022). Recent advances in lignocellulosic biomass for biofuels and value-added bioproducts - A critical review. *Bioresource Technology*, *344*(Part B), Article 126195. https://doi.org/10.1016/j.biortech.2021.126195

Azwanida, N. N. (2015). A review on the extraction methods use in medicinal plants, principle, strength and limitation. *Medicinal and Aromatic Plants*, *4*(3), Article 196. https://doi.org/doi:10.4172/2167-0412.1000196

Boundy-Mills, K., Karuna, N., Garay, L. A., Lopez, J. M., Yee, C., Hitomi, A., Nishi, A. K., Enriquez, L. L., Roberts, C., Block, D. E., & Jeoh, T. (2019). Conversion of cassava leaf to bioavailable, high-protein yeast cell biomass. *Journal of the Science of Food and Agriculture*, *99*(6), 3034–3044. https://doi.org/10.1002/jsfa.9517

Box, G. E. P., & Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(1), 1–38. https://doi.org/10.1111/j.2517-6161.1951.tb00067.x

Cavazzuti, M. (2013). Design of experiments. In M. Cavazzuti (Ed.), *Optimization methods: From theory to design scientific and technological aspects in mechanics* (pp. 13–42). Springer. https://doi.org/10.1007/978-3-642-31187-1_2

Chahyadi, A., & Elfahmi. (2020). The influence of extraction methods on rutin yield of cassava leaves (*Manihot esculenta* Crantz). *Saudi Pharmaceutical Journal*, *28*(11), 1466–1473. https://doi.org/10.1016/j.jsps.2020.09.012

Chaiareekitwat, S., Latif, S., Mahayothee, B., Khuwijitjaru, P., Nagle, M., Amawan, S., & Müller, J. (2022). Protein composition, chlorophyll, carotenoids, and cyanide content of cassava leaves (*Manihot esculenta* Crantz) as influenced by cultivar, plant age, and leaf position. *Food Chemistry*, *372*, Article 131173. https://doi.org/10.1016/j.foodchem.2021.131173

Coşgun, A., Günay, M. E., & Yıldırım, R. (2021). Exploring the critical factors of algal biomass and lipid production for renewable fuel production by machine learning. *Renewable Energy*, *163*, 1299–1317. https://doi.org/10.1016/j.renene.2020.09.034

Díez Valbuena, G., García Tuero, A., Díez, J., Rodríguez, E., & Hernández Battez, A. (2024). Application of machine learning techniques to predict biodiesel iodine value. *Energy*, *292*, Article 130638. https://doi.org/10.1016/j.energy.2024.130638

Hue, K. T., Van, D. T. T., Ledin, I., Wredle, E., & Spörndly, E. (2012). Effect of harvesting frequency, variety and leaf maturity on nutrient composition, hydrogen cyanide content and cassava foliage yield. *Asian-Australasian Journal of Animal Sciences*, *25*(12), 1691–1700. https://doi.org/10.5713/ajas.2012.12052

Iweka, S. C., Ozioko, F. C., Edafiadhe, E. D., & Adepoju, T. F. (2023). Bio-oil production from ripe pawpaw seeds and its optimal output: Box-Behnken Design and Machine Learning approach. *Scientific African*, *21*, Article e01826. https://doi.org/10.1016/j.sciaf.2023.e01826

Jiang, F., Huo, L., Chen, D., Cao, L., Zhao, R., Li, Y., & Guo, T. (2023). The controlling factors and prediction model of pore structure in global shale sediments based on random forest machine learning. *Earth-Science Reviews*, *241*, Article 104442. https://doi.org/10.1016/j.earscirev.2023.104442

Jönsson, L. J., & Martín, C. (2016). Pretreatment of lignocellulose: Formation of inhibitory by-products and strategies for minimizing their effects. *Bioresource Technology*, *199*, 103–112. https://doi.org/10.1016/j.biortech.2015.10.009

Kammoun, M., Margellou, A., Toteva, V. B., Aladjadjiyan, A., Sousa, A. F., Luis, S. V., Garcia-Verdugo, E., Triantafyllidis, K. S., & Richel, A. (2023). The key role of pretreatment for the one-step and multi-step conversions of European lignocellulosic materials into furan compounds. *RSC Advances*, *13*(31), 21395–21420. http://doi.org/10.1039/d3ra01533e

Karuna, N., Buapho, P., Sukphan, S., Bootrumka, P., Poolthong, T., Kiatkittipong, W., & Jaturapiree, P. (2025). Cassava leaf extract for enhanced biobutanol production from sugarcane bagasse using *Clostridium beijerinckii*. *Biomass Conversion and Biorefinery*, *15*(14), 21247–21259. http://doi.org/10.1007/s13399-025-06649-8

Karuna, N., Jindapang, P., Saengphenchan, R., Panpedthan, J., & Supasorn, S. (2023). Cassava leaves as an alternative nitrogen source for ethanol fermentation. *Bioenergy Research*, *16*(2), 835–842. https://doi.org/10.1007/s12155-022-10473-7

Karuna, N., Zhang, L., Walton, J. H., Couturier, M., Oztop, M. H., Master, E. R., McCarthy, M. J., & Jeoh, T. (2014). The impact of alkali pretreatment and post-pretreatment conditioning on the surface properties of rice straw affecting cellulose accessibility to cellulases. *Bioresource Technology*, *167*, 232–240. http://doi.org/ 10.1016/j.biortech.2014.05.122

Katongtung, T., Phromphithak, S., Onsree, T., & Tippayawong, N. (2024). Machine learning approach for predicting hydrothermal liquefaction of lignocellulosic biomass. *Bioenergy Research*, *7*(4), 2246–2258. https://doi.org/10.1007/s12155-024-10773-0

Kim, J. S., Lee, Y. Y., & Kim, T. H. (2016). A review on alkaline pretreatment technology for bioconversion of lignocellulosic biomass. *Bioresource Technology*, *199*, 42–48. https://doi.org/10.1016/j.biortech.2015.08.085

Kok, Z. H., Mohamed Shariff, A. R., Alfatni, M. S. M., & Khairunniza-Bejo, S. (2021). Support vector machine in precision agriculture: A review. *Computers and Electronics in Agriculture*, *191*, Article 106546. https://doi.org/10.1016/j.compag.2021.106546

La Fé-Perdomo, I., Ramos-Grez, J. A., Jeria, I., Guerra, C., & Barrionuevo, G. O. (2022). Comparative analysis and experimental validation of statistical and machine learning-based regressors for modeling the surface roughness and mechanical properties of 316L stainless steel specimens produced by selective laser melting.

Journal of Manufacturing Processes, *80*, 666–682. https://doi.org/10.1016/j.jmapro.2022.06.021

Lamidi, S., Olaleye, N., Bankole, Y., Obalola, A., Aribike, E., & Adigun, I. (2022). Applications of response surface methodology (RSM) in product design, development, and process optimization. In P. Kayarogannam (Ed.), *Response surface methodology - Research advances and applications* (pp. 1–19). IntechOpen. https://doi.org/10.5772/intechopen.106763

Lammens, T. M., Franssen, M. C. R., Scott, E. L., & Sanders, J. P. M. (2012). Availability of protein-derived amino acids as feedstock for the production of bio-based chemicals. *Biomass and Bioenergy*, *44*, 168–181. https://doi.org/10.1016/j.biombioe.2012.04.021

Malek, S., Hui, C., Aziida, N., Cheen, S., Toh, S., & Milow, P. (2019). Ecosystem monitoring through predictive modeling. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (Vol. 3, pp. 1–8). Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20060-5

Manach, C., Scalbert, A., Morand, C., Rémésy, C., & Jiménez, L. (2004). Polyphenols: Food sources and bioavailability. *The American Journal of Clinical Nutrition*, *79*(5), 727–747. https://doi.org/10.1093/ajcn/79.5.727

Manoharan, A., Begam, K. M., Aparow, V. R., & Sooriamoorthy, D. (2022). Artificial Neural Networks, Gradient Boosting and Support Vector Machines for electric vehicle battery state estimation: A review. *Journal of Energy Storage*, *55*(Part A), Article 105384. https://doi.org/10.1016/j.est.2022.105384

Moldovan, M. L., Iurian, S., Puscas, C., Silaghi-Dumitrescu, R., Hanganu, D., Bogdan, C., Vlase, L., Oniga, I., & Benedec, D. (2019). A design of experiments strategy to enhance the recovery of polyphenolic compounds from *Vitis vinifera* by-products through heat reflux extraction. *Biomolecules*, *9*(10), Article 529. https://doi.org/10.3390/biom9100529

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In O. A. Montesinos López, A. Montesinos López, & J. Crossa (Eds.), *Multivariate statistical machine learning methods for genomic prediction* (pp. 109–139). Springer. https://doi.org/10.1007/978-3-030-89010-0_4

Mota, F. L., Queimada, A. J., Pinho, S. P., & Macedo, E. A. (2008). Aqueous solubility of some natural phenolic compounds. *Industrial & Engineering Chemistry Research*, *47*(15), 5182–5189. https://doi.org/10.1021/ie071452o

Öğütcü, M., Dincer Albayrak, E., & Toklucu, A. K. (2024). Optimization of organogels prepared with turpentine oil and wax mixtures via response surface methodology and determination of vaporization kinetic parameters. *Journal of the Science of Food and Agriculture*, *104*(11), 6431–6438. https://doi.org/10.1002/jsfa.13466

Olawuni, O. A., Sadare, O. O., & Moothi, K. (2024). Optimization of liquid hot water pretreatment for extraction of nanocellulose crystal from South African waste corncobs. *Chemical Engineering Communications*, *211*(1), 26–39. https://doi.org/10.1080/00986445.2023.2218269

Patil, S. S., Deshannavar, U. B., Gadekar-Shinde, S. N., Gadagi, A. H., & Kadapure, S. A. (2023). Optimization studies on batch extraction of phenolic compounds from *Azadirachta indica* using genetic algorithm and machine learning techniques. *Heliyon*, *9*(11), Article e21991. https://doi.org/10.1016/j.heliyon.2023.e21991

Petchimuthu, P., Sumanth, G. B., Kunjiappan, S., Kannan, S., Pandian, S. R. K., & Sundar, K. (2023). Green extraction and optimization of bioactive compounds from *Solanum torvum* Swartz. using ultrasound-aided solvent extraction method through RSM, ANFIS and machine learning algorithm. *Sustainable Chemistry and Pharmacy*, *36*, Article 101323. https://doi.org/10.1016/j.scp.2023.101323

Phromphithak, S., Onsree, T., & Tippayawong, N. (2021). Machine learning prediction of cellulose-rich materials from biomass pretreatment with ionic liquid solvents. *Bioresource Technology*, *323*, Article 124642. https://doi.org/10.1016/j.biortech.2020.124642

Plackett, R. L., & Burman, J. P. (1946). The design of optimal multifactorial experiments. *Biometrika*, *33(4)*, 305–325. https://doi.org/10.2307/2332195

Roy, A., & Chakraborty, S. (2023). Support vector machine in structural reliability analysis: A review. *Reliability Engineering & System Safety*, *233*, Article 109126. https://doi.org/10.1016/j.ress.2023.109126

Ruiz, H. A., Conrad, M., Sun, S.-N., Sanchez, A., Rocha, G. J. M., Romaní, A., Castro, E., Torres, A., Rodríguez-Jasso, R. M., Andrade, L. P., Smirnova, I., Sun, R.-C., & Meyer, A. S. (2020). Engineering aspects of hydrothermal pretreatment: From batch to continuous operation, scale-up and pilot reactor under biorefinery concept. *Bioresource Technology*, *299*, Article 122685. https://doi.org/10.1016/j.biortech.2019.122685

Saha, R., Chauhan, A., & Rastogi Verma, S. (2024). Machine learning: An advancement in biochemical engineering. *Biotechnology Letters*, *46*(4), 497–519. https://doi.org/10.1007/s10529-024-03499-8

Sittichoksataporn, J., & Choksuriwong, A. (2012, May 14–15). *Comparison of support vector machine's kernel function for unsmoke sheet rubber price forecasting* [Conference session]. The 10th International PSU Engineering Conference, Prince of Songkla University, Songkhla, Thailand.

SixSigma. (2024, May 31). *Screening DOE: Efficient factorial designs for identifying key variables.* https://www.6sigma.us/six-sigma-in-focus/screening-doe/

Sluiter, A., Hames, B., Ruiz, R., Scarlata, C., Sluiter, J., & Templeton, D. (2008a). *Determination of ash in biomass: Laboratory analytical procedure (LAP), Issue date: 7/17/2005* [Technical Report NREL/TP-510-42622 January 2008]. National Renewable Energy Laboratory. https://docs.nrel.gov/docs/gen/fy08/42622.pdf

Sluiter, A., Hames, B., Ruiz, R., Scarlata, C., Sluiter, J., Templeton, D., & Crocker, D. (2012). *Determination of structural carbohydrates and lignin in biomass: Laboratory analytical procedure (LAP), Issue date: 4/25/2008* [Technical Report NREL/TP-510-42618 Revised August 2012]. National Renewable Energy Laboratory. https://docs.nrel.gov/docs/gen/fy13/42618.pdf

Sluiter, A., Ruiz, R., Scarlata, C., Sluiter, J., & Templeton, D. (2008b). *Determination of extractives in biomass: Laboratory analytical procedure (LAP), Issue date: 7/17/2005* [Technical Report NREL/TP-510-42619 January 2008]. National Renewable Energy Laboratory. https://docs.nrel.gov/docs/gen/fy08/42619.pdf

Sowcharoensuk, C. (2023). *Industry outlook 2023-2025: Cassava industry*. https://www.krungsri.com/en/research

/industry/industry-outlook/agriculture/cassava/io/cassava-2023-2025

Sparg, S. G., Light, M. E., & van Staden, J. (2004). Biological activities and distribution of plant saponins. *Journal of Ethnopharmacology*, *94*(2–3), 219–243. https://doi.org/10.1016/j.jep.2004.05.016

Suriyachai, N., Weerasai, K., Upajak, S., Khongchamnan, P., Wanmolee, W., Laosiripojana, N., Champreda, V., Suwannahong, K., & Imman, S. (2020). Efficiency of catalytic liquid hot water pretreatment for conversion of corn stover to bioethanol. *ACS Omega*, *5*(46), 29872–29881. https://doi.org/10.1021/acsomega.0c04054

Thiex, N. J., Manson, H., Anderson, S., & Persson, J.-Å. (2002). Determination of crude protein in animal feed, forage, grain, and oilseeds by using block digestion with a copper catalyst and steam distillation into boric acid: Collaborative study. *Journal of AOAC International*, *85*(2), 309–317. https://doi.org/10.1093/jaoac/85.2.309

Tomás-Pejó, E., Alvira, P., Ballesteros, M., & Negro, M. J. (2011). Pretreatment technologies for lignocellulose-to-bioethanol conversion. In A. Pandey, C. Larroche, S. C. Ricke, C.-G. Dussap, & E. Gnansounou (Eds.), *Biofuels: Alternative feedstocks and conversion processes* (pp. 149–176). Academic Press. https://doi.org/10.1016/B978-0-12-385099-7.00007-3

Vinitha, N., Vasudevan, J., & Gopinath, K. P. (2023). Bioethanol production optimization through machine learning algorithm approach: Biomass characteristics, saccharification, and fermentation conditions for enzymatic hydrolysis. *Biomass Conversion and Biorefinery*, *13*(8), 7287–7299. https://doi.org/10.1007/s13399-022-03163-z

Wang, J., & Gao, R. X. (2022). Innovative smart scheduling and predictive maintenance techniques. In D. Mourtzis (Ed.), *Design and operation of production networks for mass personalization in the era of cloud technology* (pp. 181–207). Elsevier. https://doi.org/10.1016/B978-0-12-823657-4.00007-5

Wang, M. F. Z., & Fernandez-Gonzalez, R. (2017). (Machine-)Learning to analyze in vivo microscopy: Support vector machines. *Biochimica et Biophysica Acta - Proteins and Proteomics*, *1865*(11, Part B), 1719–1727. https://doi.org/10.1016/j.bbapap.2017.09.013

Widodo, A., & Yang, B.-S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, *21*(6), 2560–2574. https://doi.org/10.1016/j.ymssp.2006.12.007

Xia, Y. (2020). Correlation and association analyses in microbiome study integrating multiomics in health and disease. In J. Sun (Ed.), *Progress in molecular biology and translational science* (Vol. 171, pp. 309–491). Academic Press. https://doi.org/10.1016/bs.pmbts.2020.04.003

Yahya, A. B., Usaku, C., Daisuk, P., & Shotipruk, A. (2023). Enzymatic hydrolysis as a green alternative for glyceride removal from rice bran acid oil before γ-oryzanol recovery: Statistical process optimization. *Biocatalysis and Agricultural Biotechnology*, *50*, Article 102727. https://doi.org/10.1016/j.bcab.2023.102727

Zendehboudi, A., Baseer, M. A., & Saidur, R. (2018). Application of support vector machine models for forecasting solar and wind energy resources: A review. *Journal of Cleaner Production*, *199*, 272–285. https://doi.org/10.1016/j.jclepro.2018.07.164

Zhang, B., Liu, X., & Bao, J. (2023). High solids loading pretreatment: The core of lignocellulose biorefinery as an industrial technology – An overview. *Bioresource Technology*, *369*, Article 128334. https://doi.org/10.1016/j.biortech.2022.128334

Zhang, M., Cui, S. W., Cheung, P. C. K., & Wang, Q. (2007). Antitumor polysaccharides from mushrooms: A review on their isolation process, structural characteristics and antitumor activity. *Trends in Food Science & Technology*, *18*(1), 4–19. https://doi.org/10.1016/j.tifs.2006.07.013

Zhang, W., Li, P., Wang, L., Fu, X., Wan, F., Wang, Y., Shu, L., & Yong, L.-q. (2024). Prediction of the yield strength of as-cast alloys using the random forest algorithm. *Materials Today Communications*, *38*, Article 108520. https://doi.org/10.1016/j.mtcomm.2024.108520

SUPPLEMENTARY DATA

Table S1. Experimental runs and results of Plackett-Burman design with actual values (coded values)

Run	X_1	X_2	X_3	X_4	X_5	Y (mg of N/L)†
1	120 (-1)	60 (+1)	212 (+1)	150 (+1)	10 (-1)	291 ± 20
2	120 (-1)	60 (+1)	50 (-1)	150 (+1)	30 (+1)	864 ± 136
3	120 (-1)	10 (-1)	50 (-1)	50 (-1)	10 (-1)	283 ± 23
4	120 (-1)	10 (-1)	50 (-1)	150 (+1)	10 (-1)	212 ± 30
5	200 (+1)	60 (+1)	50 (-1)	50 (-1)	10 (-1)	173 ± 29
6	200 (+1)	10 (-1)	212 (+1)	150 (+1)	10 (-1)	282 ± 13
7	200 (+1)	10 (-1)	212 (+1)	150 (+1)	30 (+1)	835 ± 28
8	120 (-1)	10 (-1)	212 (+1)	50 (-1)	30 (+1)	1303 ± 72
9	200 (+1)	60 (+1)	212 (+1)	50 (-1)	10 (-1)	324 ± 27
10	200 (+1)	10 (-1)	50 (-1)	50 (-1)	30 (+1)	1029 ± 10
11	200 (+1)	60 (+1)	50 (-1)	150 (+1)	30 (+1)	647 ± 10
12	120 (-1)	60 (+1)	212 (+1)	50 (-1)	30 (+1)	897 ± 205

Note: †The ± value in the concentration column represents the standard deviation of duplicate measurements.

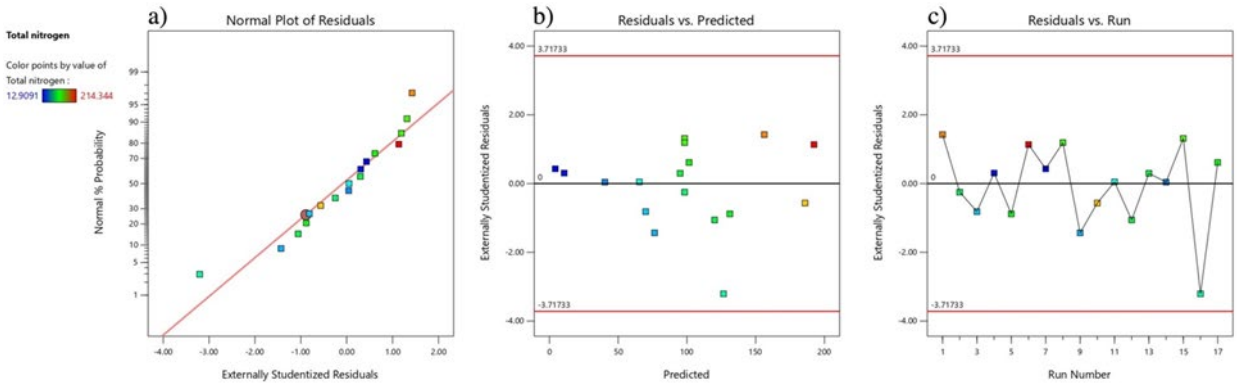

Figure S1. The plots of residuals versus (a) normal probability, (b) model predictions, and (c) run numbers from the model