



## OPEN ACCESS

## EDITED BY

Malinee Sririyanun,  
Kmutnb, Thailand

## REVIEWED BY

Neelu Raina,  
Shri Mata Vaishno Devi University, India  
Vanarat Phakeenuya,  
King Mongkut's University of Technology North  
Bangkok, Thailand

## \*CORRESPONDENCE

Yitong Niu,  
✉ itong\_niu@163.com  
Cheu Peng Leh,  
✉ cpleh@usm.my

RECEIVED 11 October 2025

REVISED 29 October 2025

ACCEPTED 07 November 2025

PUBLISHED 21 November 2025

## CITATION

Niu Y, Tye YY, Lee CK, Ahmad MI and Leh CP  
(2025) Composition-centered prediction of  
kenaf core saccharification for next-generation  
bioethanol via machine learning.  
*Front. Fuels.* 3:1722932.  
doi: 10.3389/ffuel.2025.1722932

## COPYRIGHT

© 2025 Niu, Tye, Lee, Ahmad and Leh. This is an  
open-access article distributed under the terms  
of the [Creative Commons Attribution License  
\(CC BY\)](#). The use, distribution or reproduction in  
other forums is permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication in this  
journal is cited, in accordance with accepted  
academic practice. No use, distribution or  
reproduction is permitted which does not  
comply with these terms.

# Composition-centered prediction of kenaf core saccharification for next-generation bioethanol via machine learning

Yitong Niu<sup>1\*</sup>, Ying Ying Tye<sup>1</sup>, Chee Keong Lee<sup>2</sup>,  
Mardiana Idayu Ahmad<sup>3</sup> and Cheu Peng Leh<sup>1\*</sup>

<sup>1</sup>Bioresource Technology Division, School of Industrial Technology, Universiti Sains Malaysia, Gelugor, Penang, Malaysia, <sup>2</sup>Bioprocess Technology Division, School of Industrial Technology, Universiti Sains Malaysia, Gelugor, Penang, Malaysia, <sup>3</sup>Environmental Technology Division, School of Industrial Technology, Universiti Sains Malaysia, Gelugor, Penang, Malaysia

**Introduction:** Biomass pretreatment outcomes are heterogeneous across routes and severities, and condition-centered empirical models often fail to generalize beyond the settings on which they were trained, limiting early-stage decisions about where to focus costly wet-lab effort. This study evaluates a composition-centered surrogate that treats the post-pretreatment solid composition—cellulose, hemicellulose, lignin—as the input space and predicts enzymatic glucose yield as the response for kenaf core.

**Methods:** Kenaf core solids subjected to water, dilute-acid, and alkaline pretreatments were characterized for post-pretreatment cellulose, hemicellulose, and lignin contents and hydrolyzed under a fixed enzymatic protocol to obtain glucose yield at 24 h. The curated dataset ( $n = 35$ ) was used to train Random-Forest regressors tuned by six hyperparameter optimizers (grid search, random search, Bayesian optimization, genetic algorithm, particle swarm optimization, and simulated annealing). Generalization performance was assessed using nested cross-validation and a held-out test split, with feature contributions examined via permutation importance and accumulated local effects.

**Results:** Across optimizers, held-out performance clustered tightly (test  $R^2 \approx 0.49$ – $0.55$ ; RMSE 4.42–4.69 GY%), indicating that attainable accuracy is governed more by model capacity and data coverage than by optimizer choice. Feature diagnostics converged on a cellulose-led mechanism, with cellulose showing a positive monotonic effect on yield, lignin a negative effect, and hemicellulose a weaker, context-dependent influence. Iso-yield maps in the cellulose–lignin plane delineated feasible composition windows that prioritize high-cellulose/low-lignin regions under different hemicellulose levels.

**Discussion:** Within this accuracy band, the composition-centered surrogate is best suited for uncertainty-aware screening to prune unproductive regions of composition space before targeted design-of-experiments, rather than replacing detailed process optimization. The workflow provides a transferable template for

small-sample, composition-based modeling of lignocellulosic feedstocks and can be extended to other varieties and integrated with mechanistic descriptors as data accumulate.

#### KEYWORDS

bioethanol, kenaf core, pretreatment, machine learning, yield prediction, heuristic optimization

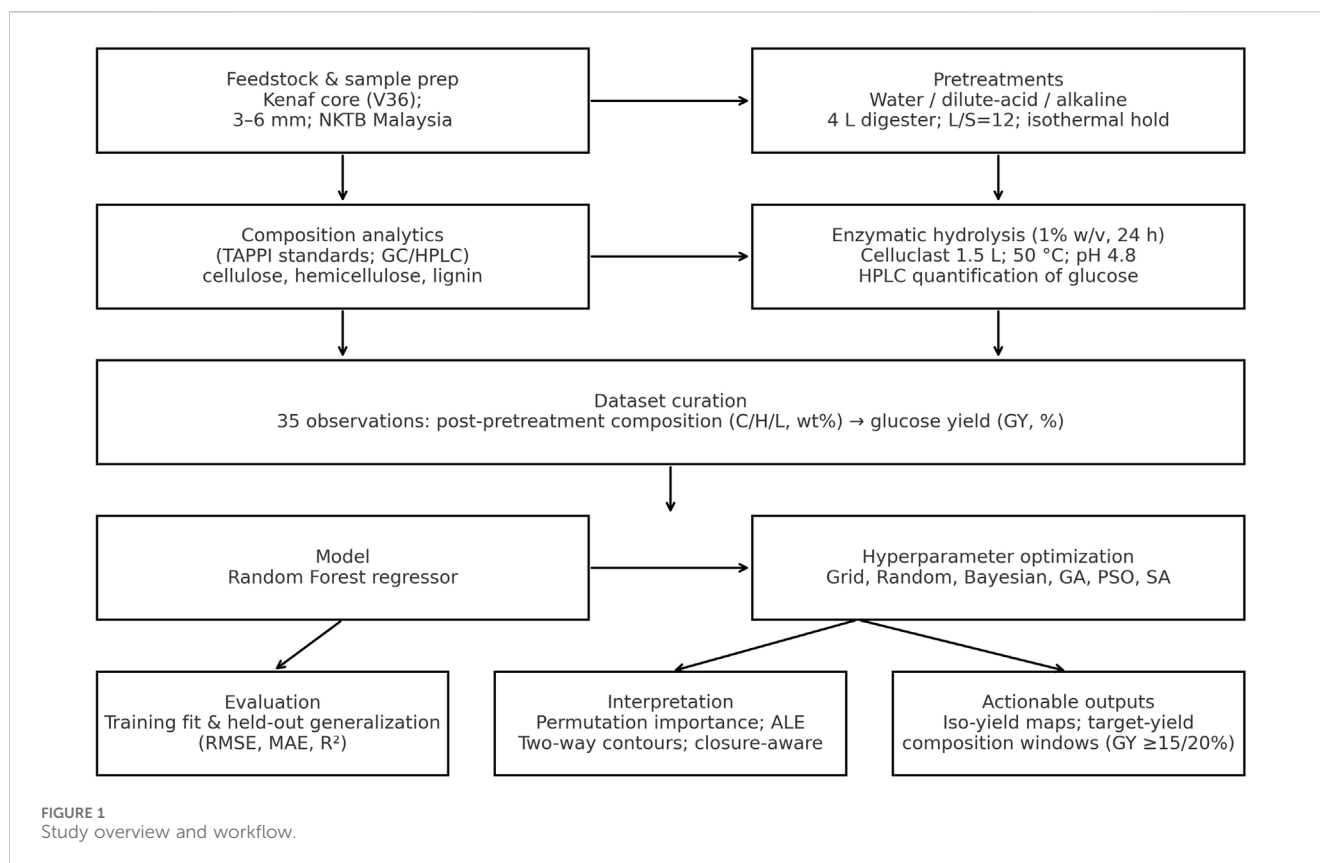
## 1 Introduction

Decarbonizing liquid fuels remains a central challenge for the energy transition, and lignocellulosic bioethanol, compatible with today's engines and infrastructure, stands out as a leading option to displace or blend with petroleum gasoline (Arias et al., 2024; Niu et al., 2024). Along the biomass-to-ethanol pathway, the pretreatment–enzymatic hydrolysis segment is widely recognized as the main performance bottleneck of the sugar platform because it concentrates the effects of biomass recalcitrance and strongly influences downstream yields and costs (Himmel et al., 2007; Yang and Wyman, 2008). The physicochemical state of the pretreated solids, especially the proportions and distributions of cellulose, hemicellulose, and lignin, governs enzyme accessibility and non-productive adsorption on lignin, thereby controlling fermentable-sugar release (Öhgren et al., 2007; Yuan et al., 2021b). A modeling strategy that takes post-pretreatment composition as the central input is therefore attractive: it abstracts away route-specific operating details while retaining the structural determinants most directly tied to hydrolysis performance, a premise supported by empirical and machine-learning studies that predict sugar yields or digestibility from composition/structure features across feedstocks and pretreatments (Namboonlue et al., 2025; Niu et al., 2025b; Niu et al., 2025a; Xie and Fan, 2025).

Kenaf (*Hibiscus cannabinus*) is a fast-growing, non-food fiber crop well adapted to warm tropical and subtropical climates, making it an attractive feedstock for bioconversion in these regions (Austin et al., 2024). Within the stem, the woody core accounts for roughly 70% of the mass (vs. ~30% bast), providing a steady and abundant process stream for valorization (Tajuddin et al., 2016). Yet the core exhibits substantial variability in chemical composition, and thus recalcitrance, across sources and processing histories, with reported ranges and location/processing effects on cellulose, hemicellulose and lignin contents (Abdullah et al., 2020). Different pretreatment routes and severities drive composition in divergent directions: alkaline methods favor delignification, whereas hydrothermal or hot-water or dilute-acid treatments primarily solubilize hemicellulose (with partial lignin relocation), so the *same nominal biomass* can yield markedly different enzymatic responses after pretreatment (Li et al., 2017; Li et al., 2022). For kenaf core specifically, side-by-side studies show that water, dilute-acid, and alkali pretreatments produce distinct post-pretreatment compositions and glucose yields (e.g., water > acid > alkali under one optimized set of conditions), and irradiation-assisted alkali can achieve high total sugar release (Jeun et al., 2015; Tye et al., 2016). In this context, predicting hydrolysis yield directly from the observed post-pretreatment composition offers a portable, route-agnostic surrogate for process screening and optimization.

Conventional optimization in biomass processing typically treats operating conditions—acid/alkali concentration, temperature, time, and liquor-to-solid ratio—as inputs and yield as the response, using design-of-experiments (DoE) and response-surface methodology (RSM). Prevailing formulations encode operating conditions as predictors, yet a state-variable formulation based on the post-pretreatment composition of the recovered solid—cellulose, hemicellulose, and lignin—offers a route-agnostic input space that is better aligned with transfer across facilities and pretreatment pathways (Kengpol and Klaiklueng, 2024; Cihan, 2025; Kumar and Chinnasamy, 2025; Saju et al., 2025). This condition-centric approach is ubiquitous in pretreatment/hydrolysis studies and reviews (Ramaraj and Unpaprom, 2019; Sharma and Sharma, 2024; Tabish et al., 2024). Sometimes these studies are augmented with semi-mechanistic or mechanistic kinetic models for cellulose/hemicellulose conversion to add physical interpretability (Jeoh et al., 2017; Yuan et al., 2021a). While powerful for *local* optimization within a single route or facility, RSM-style polynomials are deliberately local approximations and often fail to transfer across different equipment scales or pretreatment chemistries; moreover, pretreatment strategies are highly feedstock- and route-specific, limiting cross-setting generalization (Baruah et al., 2018). Compounding this, many DoE campaigns are necessarily small, which makes unbiased generalization assessment difficult without rigorous validation; classical results show that model-selection on small samples can yield optimistically biased errors unless nested cross-validation or similar safeguards are used (Hawkins, 2004; Varma and Simon, 2006). A further complication is that post-pretreatment composition data are closed (parts sum to a constant), so naive use of raw percentages can induce spurious correlation and misinterpretation; Compositional Data Analysis (CoDA) addresses this with log-ratio transforms (alr/ilr/clr) (Egozcue et al., 2003; Quinn et al., 2019). There is a need for a composition-centered, small-sample modelling workflow that (i) compares hyperparameter optimization strategies transparently, (ii) reports uncertainty and robustness, and (iii) translates predictions into actionable composition windows relevant to process decisions.

This study assemble a standardized, composition-centered dataset for kenaf core and evaluate a Random-Forest baseline tuned by six hyperparameter optimizers. The study asks whether post-pretreatment composition predicts glucose yield in a small-*n* setting, compares accuracy–efficiency across optimizers, quantifies feature effects for mechanistic consistency, delineates composition windows for target yields with uncertainty, and probes robustness to splits and representations. Interpretation relies on model-agnostic permutation importance, ALE curves, and iso-yield contouring confined to the observed convex hull, with bootstrap procedures to quantify uncertainty. This composition-centered approach provides a



route-agnostic surrogate that links measurable post-pretreatment states to expected hydrolysis performance, enabling rapid screening and go/no-go decisions without exhaustively enumerating process conditions. More broadly, it offers a reproducible template for small-sample, uncertainty-aware modeling of lignocellulosic feedstocks that can be extended to other varieties and integrated with mechanistic descriptors as data accumulate.

## 2 Methods

The study proceeds in two parts: (i) experimental acquisition of post-pretreatment composition and HPLC-measured glucose yield, and (ii) machine-learning modeling on the composition triplet using a Random-Forest tuned by six optimizers. [Figure 1](#) summarizes the end-to-end workflow.

### 2.1 Experimental data acquisition

#### 2.1.1 Feedstock and sample preparation

Kenaf (*Hibiscus cannabinus*) core of variety V36 was supplied as ground material by the National Kenaf and Tobacco Board (NKTB, Malaysia). The as-received particles had a nominal length of 3–6 mm and represented the core fraction (not bast fibers). Upon receipt, the material was homogenized by gentle tumbling, and visible foreign matter was removed manually.

#### 2.1.2 Pretreatments

Water, acid, and alkaline pretreatments of kenaf core fibre were performed in a 4-L stationary stainless-steel digester (NAC Autoclave Co., Ltd., Japan) equipped with a microcomputer-controlled thermocouple. A constant liquor-to-solid ratio of 12:1 mL g<sup>-1</sup> was used in all experiments. After the reactor reached the set temperature, the reaction was held isothermally for the prescribed time. At the end of each run, the liquor was drained and the solids were washed with tap water to neutral pH, spin-dried to remove surface liquor, and stored refrigerated for subsequent analyses. The pretreated-solid yield (wt%) was calculated on an oven-dry basis as the ratio of oven-dried mass after pretreatment to the oven-dried mass of the raw feedstock.

The range of pre-processing conditions was determined through previous related research, and the range that had been optimized was selected ([Tye et al., 2013](#); [Ying TYE et al., 2017](#)): hot water 150 °C–180 °C for 30–60 min; dilute acid 120 °C–140 °C for 45–90 min at 1.0%–2.0% (v/v) H<sub>2</sub>SO<sub>4</sub>; alkaline 100 °C–140 °C for 45–60 min at 1.0%–3.0% (w/v) NaOH.

#### 2.1.3 Composition analysis

Air-dried samples of both untreated and pretreated kenaf core were milled using an IKA® MF 10 basic microfine grinder and sieved to pass 2.0 mm prior to chemical analysis. For the untreated biomass, extractives were determined according to TAPPI 204 cm-97 (ethanol–toluene extraction, slight modifications). Following the pretreatment operations, the solids were considered extractive-free, and extractives were not quantified because

extractives co-precipitation with lignin during the assay can bias the measurement (Burkhardt et al., 2013).

Lignin and carbohydrate fractions were then quantified on the extractive-free basis. Klason lignin was measured following TAPPI 222 om-02, and acid-soluble lignin according to TAPPI UM 250 (Sluiter et al., 2008). Holocellulose and its fractions— $\alpha$ -,  $\beta$ -, and  $\gamma$ -cellulose—were determined using the procedure of Álvarez et al. (2018) together with JIS 8101. Carbohydrate composition was analyzed per TAPPI 249 cm-00 by gas chromatography (GC) of alditol-acetate derivatives using an FID and a DB-225 column (30 m  $\times$  0.25 mm  $\times$  0.25  $\mu$ m; J&W Scientific, Folsom, CA, United States of America). The GC oven was held isothermally at 220 °C for 30 min; helium served as the carrier gas at 25 mL min<sup>-1</sup> with a 50:1 split, and 1  $\mu$ L injections.

All measurements were performed in triplicate and are reported as mean  $\pm$  standard deviation. Unless otherwise stated, mass fractions are given on a dry, extractive-free basis for untreated material and on a dry basis for pretreated solids. Mass-closure checks were conducted to confirm internal consistency between lignin and carbohydrate fractions.

### 2.1.4 Enzymatic hydrolysis

For both the preliminary and optimization stages, enzymatic saccharification was performed with Celluclast<sup>®</sup> 1.5 L (Novozymes A/S, Denmark; 70 FPU mL<sup>-1</sup>). Reactions were conducted in 250-mL Erlenmeyer flasks containing the pretreated (or untreated) kenaf core solids suspended in 0.05 M citrate buffer (pH 4.8) at a substrate concentration of 1% (w/v; 10 g L<sup>-1</sup>). After adding the enzyme (1 mL Celluclast per flask) the flasks were tightly sealed and incubated in a shaking water bath at 50 °C and 120 rpm for 24 h.

At the end of incubation, the flasks were boiled for 10 min to terminate the reaction, and the slurries were centrifuged at 4,000 g for 10 min to remove unhydrolyzed residues. The supernatant was stored at -4 °C until analysis. Reducing sugar concentrations and yields were quantified by HPLC, using D-glucose as the calibration standard; results are reported as glucose equivalents. All assays were carried out in triplicate, and data are expressed as mean  $\pm$  standard deviation.

### 2.1.5 HPLC analytics and yield definition

The concentrations of individual reducing sugars in the supernatants of the enzymatic hydrolysates were determined by high-performance liquid chromatography (HPLC). The HPLC system (Agilent 385-ELSD) was equipped with a Hi-Plex Ca column (300  $\times$  7.7 mm). Distilled-deionized water served as the eluent at a flow rate of 0.6 mL min<sup>-1</sup>. Prior to injection, supernatants were passed through a 0.22  $\mu$ m syringe filter, and 20  $\mu$ L of the enzyme-free sample was injected. Sugar concentrations were calculated from calibration curves constructed with D-glucose standards. All measurements were performed in triplicate, and values are reported as mean  $\pm$  standard deviation.

The enzymatic saccharification yield on a volumetric basis was defined as:

$$\text{Enzymatic saccharification yield (\%)} = \frac{C_{\text{glucose,HPLC}} (\text{g L}^{-1})}{C_{\text{substrate,initial}} (\text{g L}^{-1})} \times 100$$

where  $C_{\text{substrate,initial}} = 10 \text{ g L}^{-1}$  for all hydrolysis runs (1% w/v). These HPLC-based yields were used consistently in subsequent data analysis and model development.

## 2.2 Machine-learning modeling

### 2.2.1 Dataset curation, problem setup and features

Measurements were aggregated into a curated dataset containing post-pretreatment cellulose, hemicellulose, lignin (wt %, dry basis) and HPLC-measured glucose yield (%) at 1% (w/v) after 24 h hydrolysis. Untreated composition was reported on a dry, extractive-free basis; pretreated solids on a dry basis. Basic QA (range/units harmonization, mass-closure, missing/duplicate checks) was applied; no records were removed, yielding  $n = 35$  (water/acid/alkali; GY 0.40%–25.50%).

We formulate a supervised regression problem to learn  $y$  (glucose yield, %) from  $X$  comprising C/H/L; pretreatment type (water/acid/alkali) was one-hot encoded only in sensitivity analyses. No imputation or scaling was required for the main models; engineered ratios were explored but not used in primary results to maximize portability.

A fixed 28/7 train–test split (no shuffling) was used for like-for-like comparison and to avoid perturbing potential batch/order effects in a small test set. Tuning used training data only; outer stratified  $k$ -fold CV (mean  $\pm$  SD) is the primary generalization estimate, and repeated shuffled hold-outs in the Supplement yield a comparable accuracy band.

### 2.2.2 Model

We employed a Random-Forest regressor as the primary model. The forest consists of  $T$  regression trees trained on bootstrap samples; at each split a random subset of features is considered (parameter `max_features`). Splits are chosen to minimize squared-error impurity (variance reduction):

$$\Delta I = \text{Var}(S) - \frac{n_L}{n} \text{Var}(S_L) - \frac{n_R}{n} \text{Var}(S_R)$$

and the prediction is the arithmetic mean of tree outputs:

$$\hat{y}(x) = \frac{1}{T} \sum_{t=1}^T h_t(x)$$

To ensure small-sample robustness, tree complexity is controlled via `max_depth`, `min_samples_split`, `min_samples_leaf`, `min_impurity_decrease`, and (optionally) `max_leaf_nodes`; `bootstrap = True` and out-of-bag scoring was not used. When included in sensitivity analyses, pretreatment type is one-hot encoded; numerical features (C/H/L) are used without scaling.

### 2.2.3 Hyperparameter optimization

Random-forest hyperparameters were tuned within a common search space and selected by minimizing three-fold cross-validated mean-squared error computed on the training set; the held-out test set was consulted once after selection. Because several hyperparameters are integer-valued, proposals from continuous optimizers were projected onto the feasible set and rounded prior to evaluation. To avoid degenerate forests in a small-sample regime, capacity safeguards were enforced (minimum depth, upper bounds

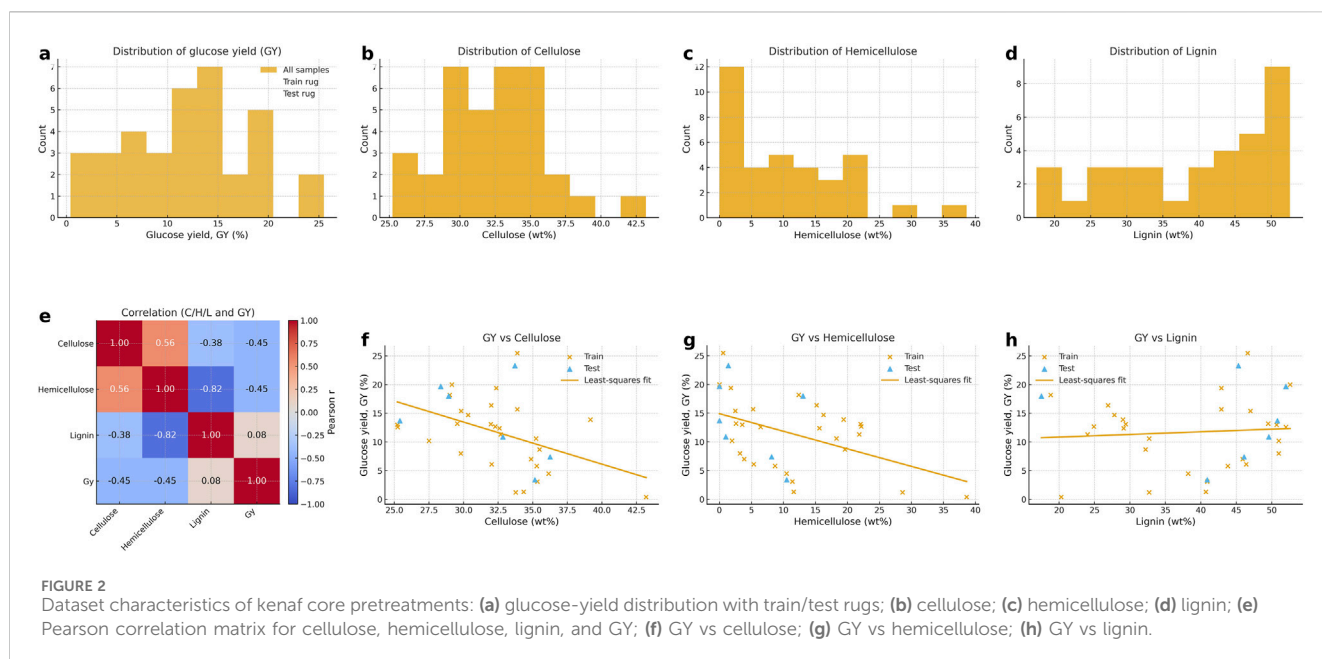


TABLE 1 Descriptive statistics of the kenaf dataset.

Split	Variable	n	Mean	SD	Min	Q1	Median	Q3	Max	Skewness
Train	Cellulose	28	32.58	3.86	25.22	29.82	32.48	34.97	43.19	0.34
Train	Hemicellulose	28	11.76	9.59	0	3.47	10.94	18.56	38.61	0.8
Train	Lignin	28	38.21	10.59	18.8	29.08	40.8	46.7	52.6	-0.23
Train	Gy	28	11.23	6.13	0.4	6.78	12.5	14.88	25.5	0.05
Test	Cellulose	7	31.52	4.01	25.4	28.64	32.84	34.43	36.25	-0.25
Test	Hemicellulose	7	4.87	5.54	0	0.5	1.38	9.32	13.05	0.36
Test	Lignin	7	43.17	11.94	17.5	43.1	46.1	50.2	52	-1.3
Test	Gy	7	13.77	7.07	3.4	9.15	13.7	18.85	23.3	-0.1
All	Cellulose	35	32.36	3.85	25.22	29.67	32.59	35.02	43.19	0.22
All	Hemicellulose	35	10.38	9.29	0	2.54	8.69	15.94	38.61	0.94
All	Lignin	35	39.2	10.87	17.5	29.25	42.9	48.25	52.6	-0.48
All	Gy	35	11.74	6.3	0.4	7.2	12.6	15.55	25.5	0.05

on leaf and split fractions, and an impurity-decrease gate). All optimizers operated under the same data split and evaluation budget.

Grid search evaluated a deterministic lattice and served as a transparent benchmark for efficiency and stability. Random search sampled from the priors and is appropriate when a small number of capacity-controlling knobs dominate performance. Bayesian optimization used a surrogate-based acquisition strategy to balance exploration and exploitation under noisy cross-validation objectives. Genetic algorithm, particle swarm optimization, and simulated annealing explored the space via population-based recombination or probabilistic hill-climbing with feasibility projection along integer axes. Final settings for the common

search space, capacity safeguards, iteration budgets, and optimizer-specific controls (for example, population size and generations, swarm and inertia coefficients, annealing schedule, and Bayesian iterations/acquisition) are reported in [Supplementary Table S1](#).

### 2.2.4 Evaluation design

Generalization was estimated with nested cross-validation. The outer loop used 5 folds (approximately 28/7 per split), stratified by pretreatment type (water/acid/alkali) to preserve class proportions. For each outer split, the inner loop applied 3-fold CV on the training portion to select hyperparameters (Section 2.2.3). The model was then refit on the full outer-training set and evaluated once on the

TABLE 2 Unified performance of tuned Random-Forest models on the training split and the held-out test split.

Optimizer		MSE	RMSE	MAE	MAPE	R <sup>2</sup>
Grid search	Train	4.25	2.06	1.49	17.63	0.88
	Test	21.24	4.61	3.46	24.09	0.50
Random search	Train	4.35	2.09	1.52	18.09	0.88
	Test	21.28	4.61	3.58	25.11	0.50
Bayesian optimization	Train	3.91	1.98	1.46	17.68	0.89
	Test	22.02	4.69	3.56	25.10	0.49
Genetic algorithm	Train	11.68	3.42	2.58	28.35	0.68
	Test	19.75	4.44	3.81	28.48	0.54
Particle swarm (PSO)	Train	36.27	6.02	4.87	43.89	0.00
	Test	-	-	-	-	-
Simulated annealing	Train	16.61	4.08	3.13	31.69	0.54
	Test	19.52	4.42	3.98	31.61	0.54

outer-test fold. All optimizers used identical folds and the same inner evaluator; random seeds were fixed to ensure comparability. The historical single 28/7 train–test split is reported only for reference in the Supplementary.

Performance metrics followed common practice. The primary metrics were R<sup>2</sup> and RMSE (RMSE =  $\sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$ ); MAE was reported as a complementary scale-robust measure.

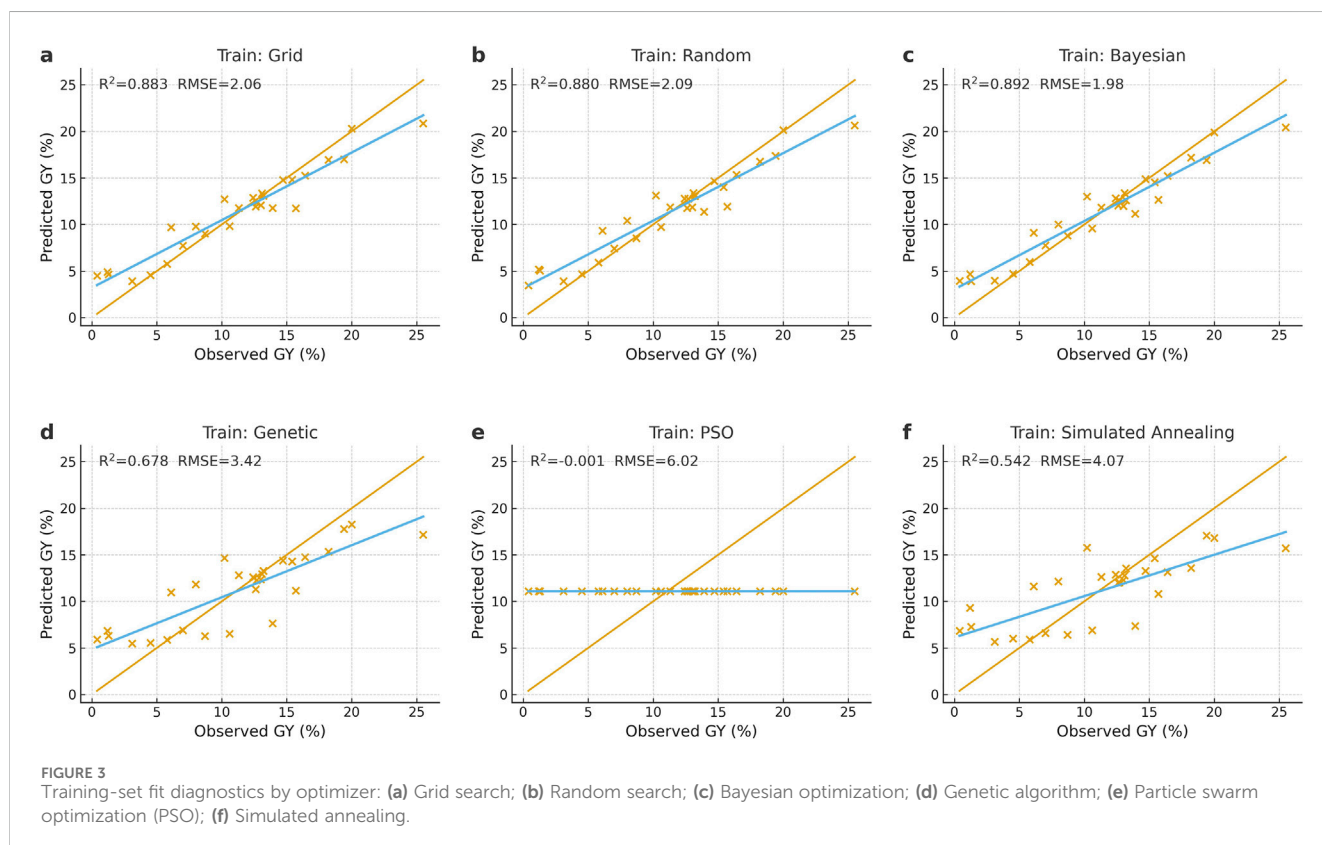
MAPE was not used as a primary metric because very low yields inflate percentage errors; where shown, a small denominator clamp was applied. Fold-wise results were aggregated as mean ± SD across the five outer folds. Residual diagnostics (residuals vs. observed yield and by pretreatment) were examined to contextualize error patterns.

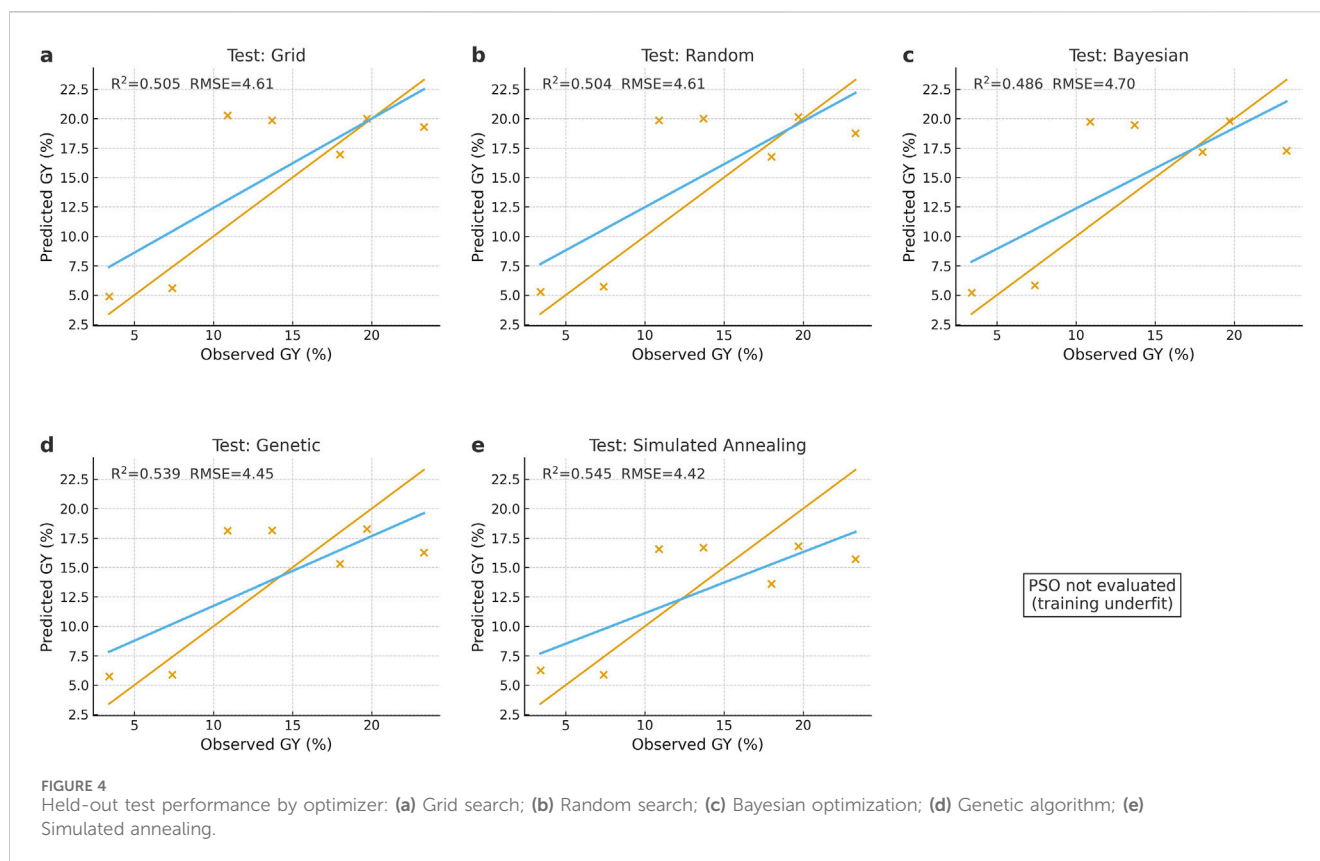
To compare optimizers fairly, wall-clock training time was recorded on the training portion of each outer split, and no test-fold information was available during tuning. All metrics were computed with standard implementations (scikit-learn), and discrete hyperparameters proposed by continuous optimizers were projected to valid values before scoring.

### 3 Result and discussion

#### 3.1 Dataset characteristics

Glucose yield (GY) denotes the experimental, mass-based glucose yield relative to the initial dry mass of the pretreated solid. As shown in Figure 2a, GY spans 0.40%–25.50% with a right-skewed distribution and a pronounced interior band around ~12–16%, flanked by sparse lower and upper tails. Train/test rugs in Figure 2a indicate comparable coverage across the split, so the held-out set samples the same interior region that dominates the data. Table 1 reports the corresponding summary statistics (counts, quartiles, dispersion, skewness), which confirm the concentration of probability mass in the interior and the rarity of extreme outcomes.





Post-pretreatment composition spans a wide range that reflects the fingerprints of the three routes. As shown in Figures 2b–d, cellulose typically falls between about 25 and 43 wt% with occasional enrichment into the low-40s; hemicellulose ranges from 0 to 39 wt% with a visible cluster near zero that signals extensive removal under dilute-acid conditions; lignin varies from roughly 18 to 53 wt%, and the upper tail marks cases with limited delignification typical of water or mild severity. Taken together, the shifts in cellulose, hemicellulose, and lignin delineate a heterogeneous composition space in which distinct pretreatment outcomes are directly observable from the distributional shapes.

Pairwise associations reflect both process co-variation and the arithmetic constraint of closed compositions. In Figure 2e, cellulose, hemicellulose, and lignin display negative correlations—most prominently the cellulose–lignin pair—that arise in part because their fractions must sum to a constant, so an increase in one component mechanically suppresses the others (Gloor et al., 2017). This closure effect cautions against causal interpretation of raw correlations. Associations with glucose yield are directionally consistent with domain expectations—higher cellulose aligns with higher yield, higher lignin aligns with lower yield—while hemicellulose shows a weaker pattern, but these signs should be read as descriptive diagnostics pending model-based attribution that corrects for correlation structure.

One-dimensional relationships in Figures 2f–h show a clear monotonic increase of glucose yield with cellulose and a monotonic decrease with lignin. The association with hemicellulose is weaker and becomes irregular near the margins where data support thins. Boundary sparsity is most evident near hemicellulose  $\approx 0\%$ , at the high-lignin end,

and in the upper tail of yield; these extremes are documented in Table 1. Predictions are therefore more reliable within the well-sampled interior of the composition domain, while estimates near the convex-hull edge involve greater uncertainty and a higher risk of local extrapolation.

### 3.2 Overall performance on the training and held-out sets

Held-out performance clusters in a narrow band: test  $R^2$  falls within 0.49–0.55 and RMSE within 4.42–4.69 GY%, as shown by the observed–predicted fits in Figure 4. Differences among optimizers are modest at this sample size, so attainable accuracy is governed chiefly by model capacity and data coverage rather than the choice of optimizer. Detailed metrics are consolidated in Table 2.

The training panels in Figure 3 separate the optimizers into two capacity regimes: grid, random, and Bayesian searches achieve tight in-sample fits, whereas GA and SA adopt more conservative settings; PSO underfits and is not carried forward. The train–test gaps reported in Table 2 mirror this split. Methods that emphasize expressiveness obtain lower training error but exhibit larger gaps on the held-out set, a pattern that is most visible near sparsely sampled boundaries in Figure 4 and is consistent with higher variance at small sample size. GA and SA reduce the gap through shallower trees and larger leaves, yet the resulting predictions show systematic bias, with under-prediction in the upper-yield tail and higher MAE despite competitive RMSE. Overall, the differences reflect distinct bias–variance profiles rather than substantive shifts in mean accuracy.

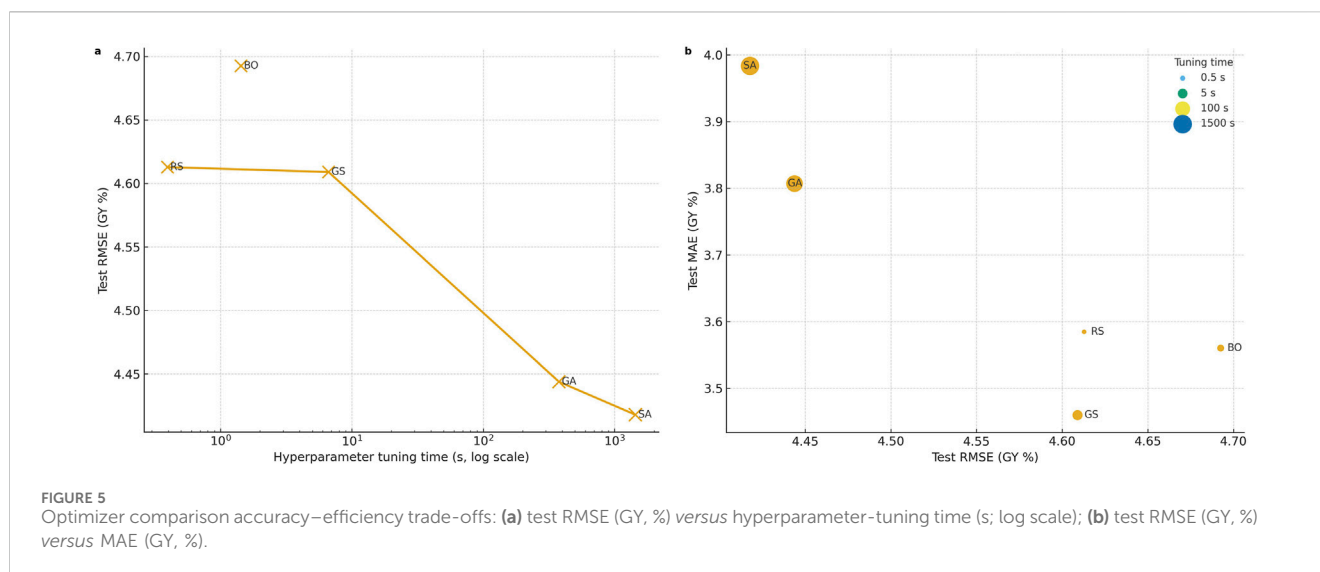


FIGURE 5 Optimizer comparison accuracy–efficiency trade-offs: (a) test RMSE (GY, %) versus hyperparameter-tuning time (s; log scale); (b) test RMSE (GY, %) versus MAE (GY, %).

The objective for each configuration is a cross-validated loss on  $n = 35$  samples, which introduces evaluation noise. Several hyperparameters are integer-valued and enforced by projection/rounding, so the search landscape is mixed-discrete and non-smooth with plateaus. Under this topology, PSO tends to collapse into over-regularized corners after projection; GA/SA favour conservative settings under noisy selection/cooling, reducing variance but introducing bias in the upper-yield tail. Random and Bayesian are more tolerant to noise and focus on a few capacity-controlling knobs, which helps explain the modest differences in mean accuracy.

Efficiency considerations favor methods that reach near-frontier accuracy with short tuning times. As shown in Figures 5a,b, random search achieves accuracy close to the best observed at a fraction of the computational cost, and Bayesian optimization performs similarly while providing guided exploration. Grid search offers no accuracy advantage relative to random search yet incurs substantially higher wall-clock time, so it is dominated. GA and SA require minute-scale budgets for a slight RMSE edge, accompanied by a higher MAE profile; they are therefore appropriate only when marginal error reductions justify the time. PSO expends considerable effort without reliable improvements in this topology. The capacity summaries in Table 3 align with these efficiency patterns and explain why methods that emphasize expressiveness are fast but variance-prone, whereas robustness-oriented configurations trade time for bias.

Within this accuracy band, method choice should be guided by the error profile and uncertainty, particularly MAE and prediction-interval width with computational cost, rather than by marginal differences in  $R^2$ . In practice, prioritizing noise-tolerant, efficient search while controlling model capacity offers the most reliable path to usable screening performance.

### 3.3 Optimizer comparison

Across the fixed split, tuned forests show tightly clustered accuracy (as reported in Table 2 and shown in Figure 4). Within

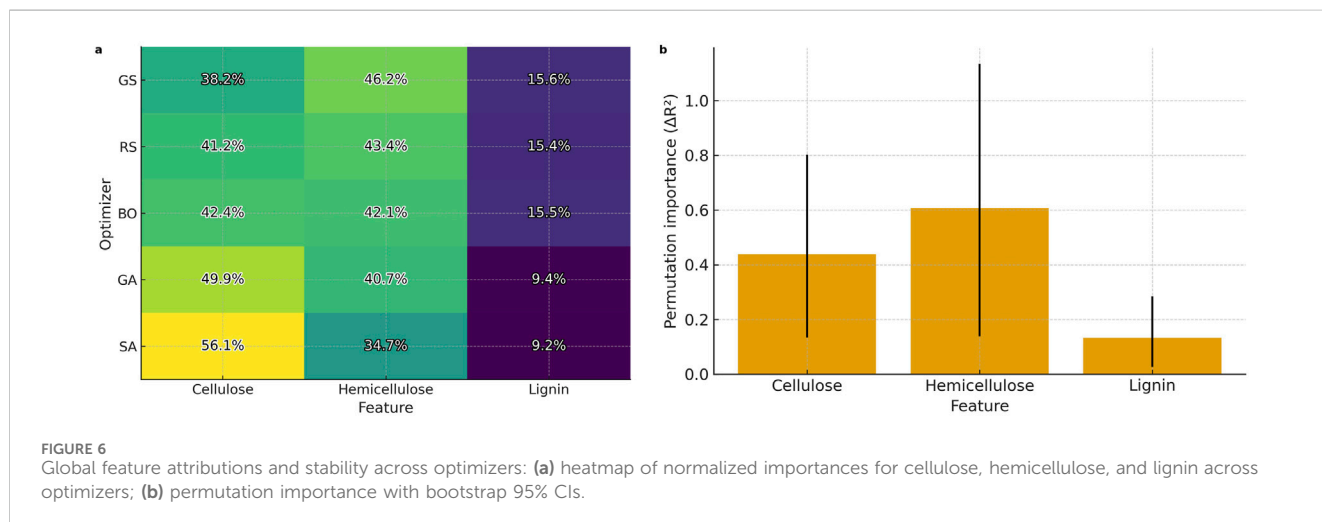
this band, differences are systematic but modest: SA/GA trade a slight RMSE edge for higher MAE; grid/random are mid-band with balanced profiles; Bayesian optimization tracks random; PSO is excluded due to training underfit (see Figure 3).

These outcomes reflect how each method navigates the mixed discrete–continuous, noisy hyperparameter space of random forests under small  $n$ . The objective for every candidate is a cross-validated loss estimate—noisy at this sample size—and feasibility is enforced by projection or rounding on integer dimensions (depth, leaf size, split thresholds, impurity gate, tree count, bootstrap). The resulting landscape is non-smooth with plateaus (Varma and Simon, 2006; Probst et al., 2019). In such settings, population-based methods that rely on momentum or recombination face two characteristic failure modes. First, PSO updates positions using velocity in a landscape where gradients are effectively undefined after discretization; combined with feasibility projection, the swarm tends to converge toward over-regularized corners, yielding near-constant predictions (Rezaee Jordehi and Jasni, 2015). Second, GA (and to a lesser extent SA) operates under minute-scale budgets with small populations; selection and cooling under noisy fitness favour conservative configurations that shrink variance and generalization gaps but can introduce systematic bias in loss profiles (Jin and Branke, 2005; Rakshit et al., 2017; Wang et al., 2004). By contrast, random search and Bayesian optimization are relatively noise-tolerant and effective when performance is governed by a few capacity-controlling knobs; they explore those axes directly without being pulled by momentum into discretization-induced attractors.

The final configurations corroborate these regimes without requiring exhaustive enumeration in the main text. Methods that prioritise expressiveness (grid, random, Bayesian) converge to deeper trees with small leaves and permissive impurity gates, delivering tight training fits at the expense of larger generalization gaps in sparsely sampled corners. Methods that prioritise robustness (GA, SA) adopt shallower trees and larger leaves, producing small gaps but higher MAE owing to under-prediction at the upper-yield tail. Full hyperparameter settings are reported in Table 3.

TABLE 3 Optimized Random-Forest hyperparameters by optimizer.

Parameter	Grid search	Random search	Bayesian optimization	Genetic algorithm	Particle swarm optimization	Simulated annealing
Training time (raw)	6.648 s	0.395 s	1.432 s	6 min17.389s	4 min38.662s	24 min0.812s
Training time (s)	6.648	0.395	1.432	377.389	278.662	1,440.812
Data split	1	1	1	1	1	1
Shuffle	No	No	No	No	No	No
Cross-validation	No	No	No	No	No	No
Split criterion	squared_error	squared_error	squared_error	squared_error	squared_error	squared_error
Max features (at split)	All features	All features	All features	All features	All features	All features
Min samples split	2	2	2	7	50	2
Min samples leaf	1	1	1	2	50	3
Min weight fraction leaf	0	0	0	0.079	0	0
Max depth	10	10	10	16	5	2
Max leaf nodes	50	50	50	45	150	42
Min impurity decrease	0	0	0	0.500	0.987	0
Number of trees (n_estimators)	100	100	100	58	50	91
Bootstrap sampling	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Out-of-bag scoring	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE



Efficiency comparisons align with these mechanisms (Figures 5a,b). Random search achieves accuracy close to the best observed at seconds-level tuning time and, together with Bayesian optimization, forms the empirical efficiency frontier (Nguyen, 2019). Grid search is dominated: similar accuracy at markedly higher wall-clock. GA and SA sit in the “slower-but-slightly-more-accurate” corner of RMSE with large MAE and minute-scale budgets. PSO incurs substantial time without reliable gains in this topology (Hyndman and Koehler, 2006).

Taken together, the results support a simple prescription for datasets of this size and structure. Random search is a pragmatic default; Bayesian optimization is a near-substitute when guided exploration is desired or prior structure is available. GA/SA are warranted only when marginal reductions in RMSE justify longer budgets and a higher MAE is acceptable in practice. PSO is not recommended for small-*n*, mixed-discrete hyperparameter spaces with noisy cross-validated objectives. The dominant levers remain capacity control and data coverage, not the choice among competent noise-tolerant optimizers.

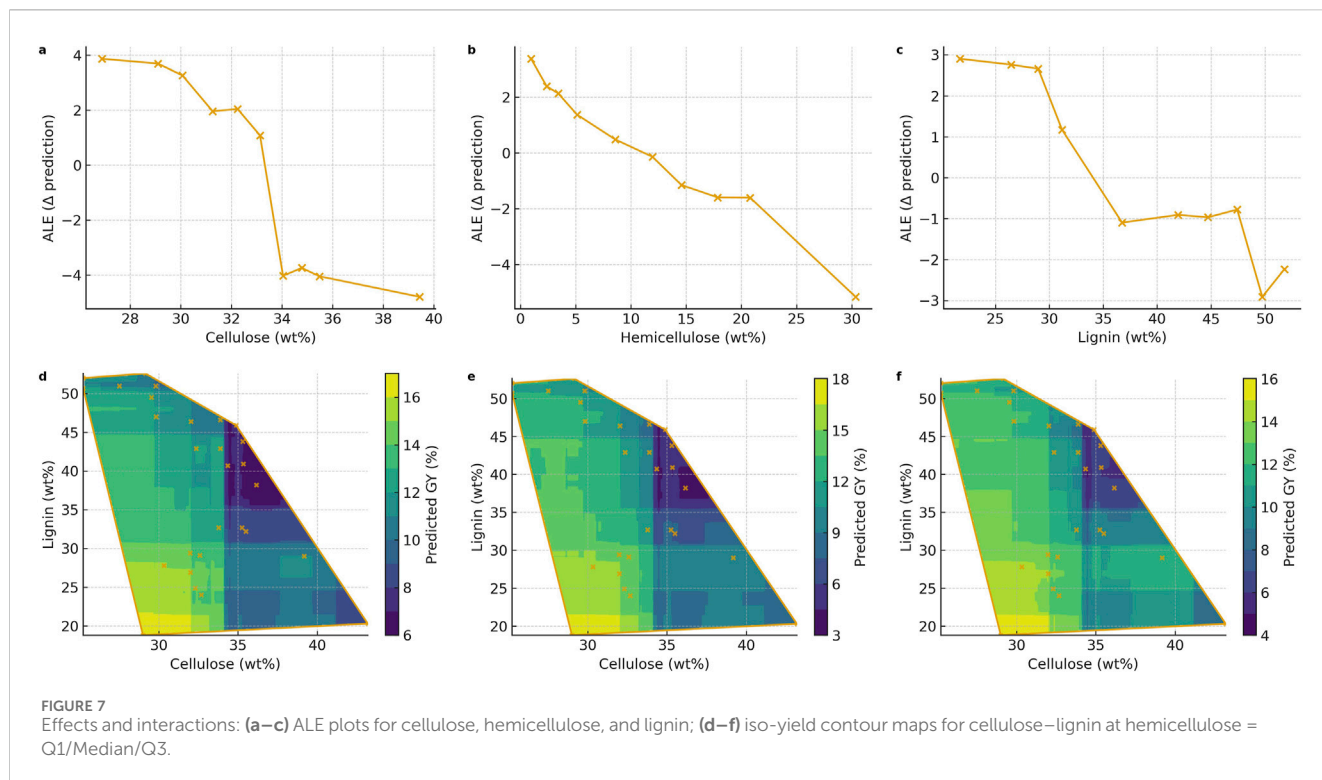


TABLE 4 Composition windows achieving target yields.

H level	Target GY (%)	Cellulose range (wt%)	Cellulose 95% CI (min)	Cellulose 95% CI (max)	Lignin range (wt%)	Lignin 95% CI (min)	Lignin 95% CI (max)	Feasible area within hull (%)	Data support (n in  H–Hfix  ≤ 2 and GY ≥ thr)
Q1 (3.47 wt %)	≥15	[28.05, 33.09]	25.69	34.51	[19.18, 27.91]	19.18	50.68	9.5	3
Q1 (3.47 wt %)	≥20	—	—	—	—	—	—	0	0
Median (10.94 wt %)	≥15	[27.84, 33.09]	26.03	34.01	[19.18, 29.81]	19.18	44.62	11.2	1
Median (10.94 wt %)	≥20	—	—	—	—	—	—	0	0
Q3 (18.56 wt %)	≥15	[28.85, 31.88]	27.91	33.03	[19.18, 21.08]	19.18	30.45	2	0
Q3 (18.56 wt %)	≥20	—	—	—	—	—	—	0	0

### 3.4 Feature contributions and mechanistic interpretation

Global attributions indicate a cellulose-led hierarchy. Figure 6a summarizes normalized importances across optimizers: lignin ranks

third for every method, while the top two features swap order—grid and random search favor hemicellulose slightly over cellulose, whereas Bayesian optimization, GA, and SA place cellulose first. This pattern indicates strong rank agreement with a single inversion between the leading pair and perfect agreement on lignin’s lower

contribution. Permutation importance with bootstrap (Figure 6b) corroborates this picture: cellulose and, to a lesser extent, hemicellulose yield the largest  $\Delta R^2$  with overlapping intervals, while lignin's  $\Delta R^2$  is markedly smaller. Taken together, attribution and ALE diagnostics support a cellulose-first ranking (cellulose more than hemicellulose; lignin negative), with hemicellulose contributing second-order variation sensitive to route and severity.

Marginal effects align with mechanistic expectations for enzymatic hydrolysis. The ALE curves in Figures 7a–c isolate one-dimensional influence while mitigating confounding from feature correlation and compositional closure. The cellulose effect is monotone increasing across its observed range, consistent with higher accessible cellulose supporting higher glucose release. Lignin displays a monotone decreasing effect, in line with its well-known role in impeding enzyme access and promoting nonproductive adsorption. Hemicellulose exhibits a weaker, less regular pattern, including shallow non-monotonicity at the margins where data are sparse. The weaker net effect is plausible: moderate hemicellulose removal can improve porosity and reduce steric hindrance, but its relationship to yield is less direct than cellulose enrichment or lignin removal and is sensitive to pretreatment pathway and severity.

Interactions between cellulose and lignin further clarify the response surface. Figures 7d–f plot iso-yield contours in the (cellulose, lignin) plane at hemicellulose fixed to the lower quartile, median, and upper quartile (values reported in Table 4), masking predictions outside the observed convex hull and overlaying the training support. Across all three slices, contours tilt toward higher cellulose and lower lignin, with the high-yield region concentrated near the high-cellulose/low-lignin corner. As hemicellulose increases from Q1 to Q3, the feasible high-yield zone contracts and shifts, indicating that elevated hemicellulose tightens the composition requirements on cellulose and lignin. The hull overlay emphasizes that the sharpest gradients occur near data boundaries; masking prevents over-interpretation of extrapolations beyond the sampled composition space.

These statistical patterns are coherent with the chemistry of the three pretreatment routes. Alkaline pretreatment drives delignification and opens cell-wall structure; the negative lignin effect on yield captured by the model reflects improved enzyme accessibility as lignin decreases. Dilute-acid pretreatment targets hemicellulose removal; the comparatively weaker and sometimes non-monotone hemicellulose signal is consistent with the idea that moderate removal aids accessibility, whereas excessive loss or accompanying side reactions can compress structure or generate inhibitors, attenuating net yield gains. Water-only or insufficiently severe conditions leave composition closer to the untreated state, aligning with the low-to-mid yield band seen in the marginal plots and distributions.

Translating these relationships into actionable operating windows highlights where composition must land to realize target yields within the observed domain. Table 4 reports feasible cellulose–lignin ranges at representative hemicellulose levels for  $GY \geq 15\%$  and  $\geq 20\%$ , computed on the convex hull with bootstrap confidence intervals and sample support counts. For  $GY \geq 15\%$ , windows exist at moderate-to-high cellulose and mid-to-low lignin across all hemicellulose slices, with nontrivial feasible area fractions. For  $GY \geq 20\%$ , the windows shrink toward the high-

cellulose/low-lignin corner and the feasible area contracts, especially at higher hemicellulose; data support also thins, signaling greater sensitivity to specification and higher uncertainty. These results argue for pretreatment strategies that simultaneously enrich cellulose and reduce lignin, while avoiding compositional corners that current data do not adequately support.

Robustness considerations temper inference from closed compositions. The principal conclusions—cellulose positive, lignin negative, hemicellulose weaker—do not rely on a particular attribution metric and persist under ALE-based visualization that is less sensitive to correlation and closure. Nevertheless, closure can bias linear associations and inflate apparent trade-offs among components.

### 3.5 Error diagnostics and subgroup analysis

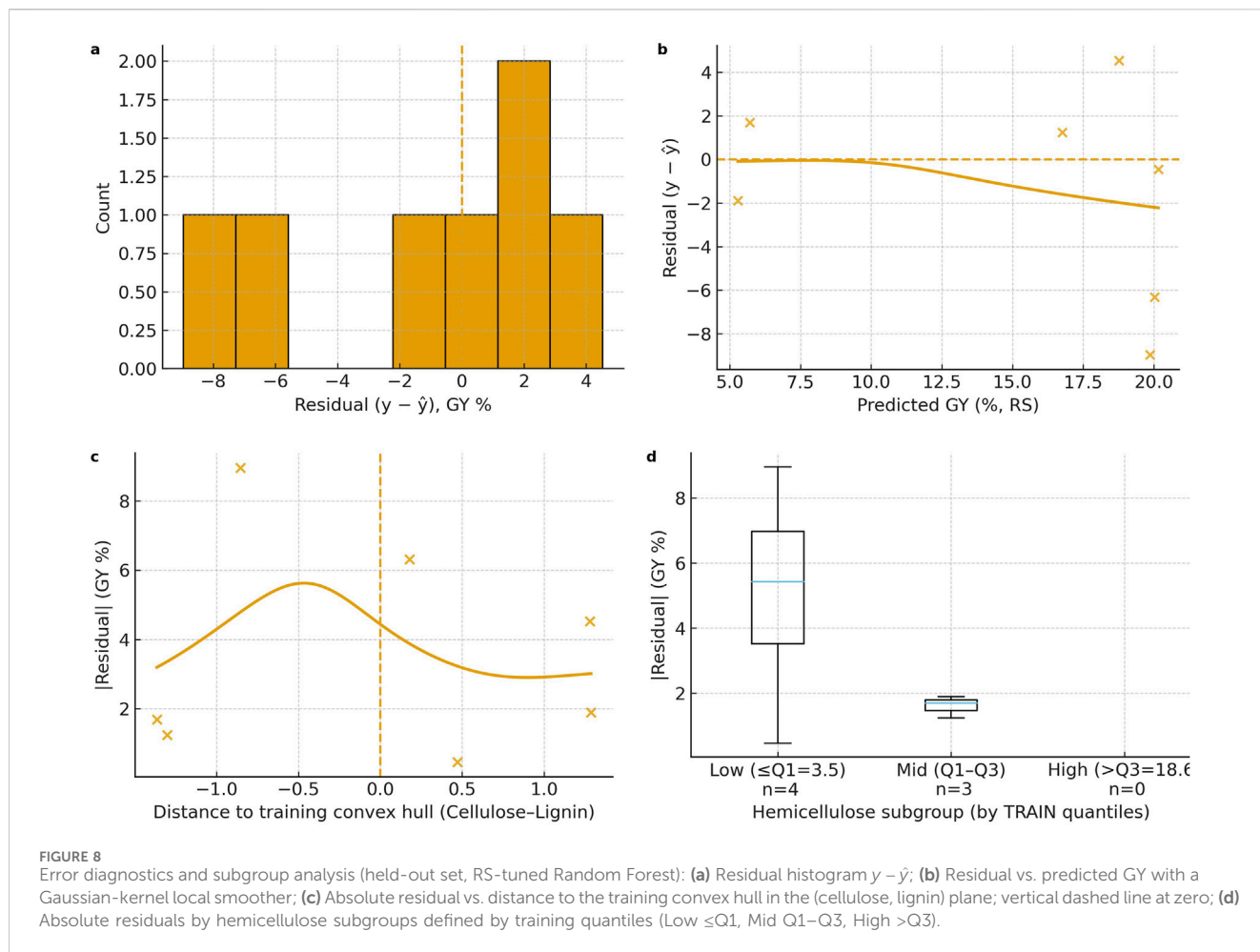
The held-out residuals  $e = y - \hat{y}$  are centered near zero but display a left-heavy tail (Figure 8a). Two large negative errors dominate the tail, yielding a small negative bias (mean  $e \sim -1.5\%$  GY) and visibly heavier tails in the Q–Q overlay, consistent with the aggregate RMSE/MAE reported in Table 2. This pattern indicates occasional over-prediction, especially for a few difficult cases.

Error magnitude is not constant across the response range. Residuals plotted against the predicted yield show a smooth downward drift at higher  $\hat{y}$  (Figure 8b), i.e., high-predicted samples are more often under-achievers on test and exhibit larger dispersion. A geometric viewpoint corroborates this: absolute residuals increase as samples approach the boundary of the training support in the (cellulose, lignin) plane, and they are largest just at or beyond the convex-hull edge (Figure 8c; negative distances denote mild extrapolation). Together these trends point to boundary sparsity and local extrapolation as primary drivers of the heaviest errors in this small-sample regime.

Stratifying by hemicellulose clarifies where the model is most reliable (Figure 8d; Table 5). Subgroups are defined by training quantiles (Q1 = 3.47 wt%, Q3 = 18.56 wt%). The mid-hemicellulose band (Q1–Q3) shows the smallest errors (RMSE  $\approx 1.6$ , MAE  $\approx 1.6\%$  GY; near-zero bias), whereas the low-hemicellulose group ( $\leq Q1$ ) exhibits much larger dispersion (RMSE  $\approx 5.9$ , MAE  $\approx 5.1\%$  GY) and a clear negative bias ( $\approx -2.8\%$  GY), reflecting several high-prediction/low-outcome cases that sit close to the composition boundary. No held-out samples fall in the high-hemicellulose group ( $> Q3$ ), underscoring the lack of coverage there and the need for caution when interpreting predictions in that slice of composition space.

Inspection of the largest-error cases supports a common mechanism: samples with very low hemicellulose and relatively high predicted yield tend to lie near the edge of the observed (cellulose, lignin) domain, where the model effectively interpolates along long, data-sparse rays. Beyond composition, unobserved factors—cellulose crystallinity and DP, pore structure, residual acetyl/ash, soluble inhibitors, batch-to-batch enzyme activity, and analytical noise—can plausibly contribute additional variance, so part of the residual is irreducible with the present feature set.

From an application standpoint, two safeguards are advisable. First, restrict use to the observed convex hull (or apply a distance-to-hull



**TABLE 5** Subgroup errors by hemicellulose level on the held-out set (RS-tuned Random Forest).

Hemicellulose subgroup	n	Bias ( $y - \hat{y}$ ), GY %	RMSE, GY %	MAE, GY %
Low ( $\leq Q1$ )	4	-2.802	5.936	5.068
Mid ( $Q1-Q3$ )	3	0.347	1.629	1.607

threshold), flagging predictions that occur near or outside the boundary where  $|e|$  inflates (Figure 8c). Second, accompany point predictions with uncertainty: quantile or conformal prediction intervals calibrated on the training split can target ~90% coverage overall and expose under-coverage in the high-yield tail. Finally, targeted data collection near the high-cellulose/low-lignin corner and in high-hemicellulose regions (absent from the test set) should reduce boundary effects and improve reliability in the very regimes that matter for process optimization.

### 3.6 Sensitivity and robustness

Robustness to resampling and split choice was first examined through repeated hold-outs and repeated k-fold evaluations under the fixed feature set. Across Monte-Carlo splits and modest train/test ratios, the dispersion of  $R^2$ , RMSE, and MAE remained within

the same performance band as the fixed split in Table 2, and the relative ordering of optimizers did not invert: search strategies that favored higher-capacity forests (grid/random/Bayesian) continued to yield tighter training fits and larger generalization gaps, whereas the more regularized configurations (GA/SA) preserved smaller gaps at a small cost in in-sample fit. Variability increased when the held-out set contained boundary or high-yield cases, but median accuracy remained stable, indicating that sampling variation, rather than optimizer idiosyncrasy, is the principal driver of spread in this small- $n$  regime.

Sensitivity to feature representation was assessed by replacing raw wt% with log-ratio coordinates for closed compositions (centered or isometric log-ratios). Under identical tuning and evaluation, performance statistics stayed in the same range, and the global attribution hierarchy—cellulose strongest positive, lignin strongest negative, hemicellulose weaker/non-monotone—was preserved. Rank agreement between importances computed in wt

% and in log-ratio space was high, and permutation-based checks gave the same ordering, suggesting that compositional closure influences magnitudes but does not reverse the substantive conclusions about directions of effect.

Randomness and local hyperparameter perturbations were then probed. Sweeps over random seeds (data shuffling, RF bootstrap, and optimizer initialization) produced overlapping error distributions, and small, coordinated nudges around the tuned settings (depth, minimum leaf/split thresholds, impurity gate, and tree count) did not materially change held-out error. These patterns are consistent with a flat optimum basin around the RS/BO solutions and with the bias–variance regimes already identified: GA/SA remain conservative across seeds, and PSO retains instability in this mixed discrete–continuous space. No single hyperparameter dominated the sensitivity; instead, capacity is governed by depth  $\times$  leaf size  $\times$  impurity threshold acting in concert.

Model-class checks supported the RF-based findings. Gradient-boosted trees or XGBoost achieved accuracy in the same band and reproduced the effect directions for cellulose, hemicellulose, and lignin when assessed by permutation importance. Simpler baselines (ridge on log-ratio features) captured the signs but underfit nonlinearities, while distance-based regressors were more brittle near data boundaries. Taken together, these contrasts indicate that the substantive conclusions are model-agnostic within common tabular learners, and that the remaining errors are more about data coverage than algorithm choice.

Finally, measurement-level robustness was evaluated by injecting realistic noise into inputs and outputs (small absolute perturbations to composition and GY consistent with analytical variability) and recomputing performance and composition windows. Aggregate metrics drifted modestly, the GY  $\geq 15\%$  windows in Table 4 were largely preserved (with slightly wider uncertainty margins), and the GY  $\geq 20\%$  windows contracted most where data are sparse—precisely the boundary zones already flagged by Figure 8. This behavior is consistent with the intuition that high-yield corners of composition space are both valuable and data-limited; small perturbations there translate into larger predictive variance.

In summary, robustness probes converge on three invariants: (i) the cellulose-positive/lignin-negative signal with hemicellulose as a weaker contributor; (ii) a capacity trade-off that explains narrow differences among optimizers on held-out accuracy; and (iii) boundary sensitivity as the dominant source of error and uncertainty. For practice, these findings motivate reporting prediction intervals, restricting use to the observed convex hull (or penalizing distance to it), and prioritizing new data in high-cellulose/low-lignin and high-hemicellulose regions to reduce sensitivity where it is currently greatest.

### 3.7 Limitations and further research

The present model attains a test-set  $R^2$  of roughly 0.5. It should therefore be interpreted as a screening surrogate that can highlight promising composition regions rather than a tool for precise point prediction at the extremes. The dataset is small ( $n = 35$ ) and concentrated in the interior of the composition domain; sparsity near high-cellulose or very low-lignin corners inflates uncertainty

and makes boundary behaviour sensitive to sampling. Several material attributes were not observed yet plausibly affect hydrolysis, such as cellulose crystallinity and degree of polymerization, pore structure, residual acetyl or ash, inhibitor burden, and batch variability in enzyme activity. Part of the residual error is likely irreducible under the current feature set. Closed-composition effects also complicate linear associations among cellulose, hemicellulose, and lignin, although attribution and accumulated local effects mitigate bias in effect directions. Finally, all windows reported here are conditional on the stated hydrolysis protocol, including solids loading, enzyme dose, temperature, pH, and duration; altering these settings will shift the response surface and requires re-calibration. External validity beyond the tested kenaf material and severity ranges remains to be established.

Future work should expand coverage where decisions are most consequential, additional observations are needed in the high-cellulose and low-lignin region and in hemicellulose-rich slices that are currently under-represented. Incorporating mechanistic descriptors—such as crystallinity index, accessible porosity, degree of polymerization, residual acetyl and ash, and soluble inhibitor metrics—may convert a share of unexplained variance into explained signal and sharpen feasible windows. Recording pretreatment severity descriptors alongside composition would help disentangle route-specific effects and support multi-objective targets that balance yield against severity. External validation across plant fractions and feedstocks should test whether the composition-centred formulation transfers without re-tuning or whether modest recalibration is required. On the methodological side, uncertainty quantification should be formalized, for example, by reporting conformal prediction intervals and by penalizing distance to the observed convex hull so that predictions near boundaries are explicitly down-weighted. Comparative baselines using alternative learners of similar capacity would help verify that the main conclusions—cellulose as the dominant positive driver, lignin as a secondary negative factor, hemicellulose as weaker and context-dependent—are model-agnostic.

## 4 Conclusion

This work introduces a composition-centered framework for predicting enzymatic glucose yield of kenaf core after pretreatment. By focusing on the measurable post-pretreatment state—cellulose, hemicellulose, and lignin—and training a Random-Forest regressor tuned with six distinct hyperparameter optimizers, the study links routine analytics to hydrolysis performance in a way that is portable across pretreatment routes and suited to small datasets.

Across optimizers, generalization converges to a narrow accuracy band, highlighting that—at the current sample size—model capacity and data coverage matter more than the specific search strategy. Random search emerges as a pragmatic default, matching the best held-out accuracy at seconds-level tuning cost, whereas more conservative settings identified by genetic algorithms or simulated annealing trade a slight RMSE advantage

for larger MAE and minute-scale runtimes. Feature attributions and ALE diagnostics are consistent with domain knowledge: cellulose exerts the strongest positive influence on yield, lignin the strongest negative, and hemicellulose a weaker, context-dependent effect. Iso-yield maps restricted to the observed domain translate these patterns into composition windows for target yields and provide immediate guidance for pretreatment targeting.

Error analysis shows that the largest uncertainties occur near the boundaries of the sampled composition space and in rare high-yield cases. For practice, we recommend guardrails that restrict use to the observed convex hull (or penalize distance to it) and the routine reporting of prediction intervals alongside point estimates to support go/no-go decisions.

The approach is subject to clear limitations: a modest sample size from a single kenaf variety and a fixed hydrolysis protocol, and the absence of structural descriptors that likely explain part of the residual variance. Future work should expand coverage, especially in the high-cellulose/low-lignin corner and at high hemicellulose, integrate mechanistic descriptors and pretreatment-severity metrics, and validate externally across varieties, feedstocks, and operating conditions. With these extensions, the composition-centered workflow outlined here can serve as a reproducible template for uncertainty-aware screening and for designing pretreatments that reliably land within composition windows linked to desired sugar yields.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Author contributions

YN: Software, Writing – original draft, Visualization, Validation, Formal Analysis, Writing – review and editing. YT: Resources, Investigation, Writing – review and editing, Data curation. CL: Supervision, Project administration, Writing – review and editing. MA: Writing – review and editing, Supervision. CL: Writing – review and editing, Supervision, Funding acquisition, Project administration.

## References

- Abdullah, H. H., Zakaria, S., Anuar, N. I. S., Mohd Salleh, K., and Syed Jaafar, S. N. (2020). Effect of harvesting time and water retting fiber processing methods on the physico-mechanical properties of kenaf fiber. *BioResources* 15, 7207–7222. doi:10.15376/biores.15.3.7207-7222
- Álvarez, A., Cachero, S., González-Sánchez, C., Montejo-Bernardo, J., Pizarro, C., and Bueno, J. L. (2018). Novel method for holocellulose analysis of non-woody biomass wastes. *Carbohydrate Polymers* 189, 250–256. doi:10.1016/j.carbpol.2018.02.043
- Arias, A., Nika, C.-E., Vasilaki, V., Feijoo, G., Moreira, M. T., and Katsou, E. (2024). Assessing the future prospects of emerging technologies for shipping and aviation biofuels: a critical review. *Renew. Sustain. Energy Rev.* 197, 114427. doi:10.1016/j.rser.2024.114427
- Austin, C. C., Mondell, C. N., Clark, D. G., and Wilkie, A. C. (2024). Kenaf: opportunities for an ancient fiber crop. *Agronomy* 14, 1542. doi:10.3390/agronomy14071542
- Baruah, J., Nath, B. K., Sharma, R., Kumar, S., Deka, R. C., Baruah, D. C., et al. (2018). Recent trends in the pretreatment of lignocellulosic biomass for value-added products. *Front. Energy Res.* 6, 141. doi:10.3389/ferg.2018.00141
- Burkhardt, S., Kumar, L., Chandra, R., and Saddler, J. (2013). How effective are traditional methods of compositional analysis in providing an accurate material balance for a range of softwood derived residues? *Biotechnol. Biofuels* 6, 90. doi:10.1186/1754-6834-6-90
- Cihan, P. (2025). Bayesian hyperparameter optimization of machine learning models for predicting biomass gasification gases. *Appl. Sci.* 15, 1018. doi:10.3390/app15031018
- Egozcue, J. J., Pawłowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Math. Geol.* 35, 279–300. doi:10.1023/A:1023818214614
- Gloor, G. B., Macklaim, J. M., Pawłowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8, 2224. doi:10.3389/fmicb.2017.02224

## Funding

The authors declare that financial support was received for the research and/or publication of this article. This work was supported by the Malaysia Ministry of Higher Education (MoHE) through the Fundamental Research Grant Scheme (FRGS) (203/PTEKIND/6711702).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/ffuel.2025.1722932/full#supplementary-material>

- Hawkins, D. M. (2004). The problem of overfitting. *J. Chem. Inf. Comput. Sci.* 44, 1–12. doi:10.1021/ci0342472
- Himmel, M. E., Ding, S.-Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W., et al. (2007). Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* 315, 804–807. doi:10.1126/science.1137016
- Hyndman, R. J., and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *Int. J. Forecast.* 22, 679–688. doi:10.1016/j.ijforecast.2006.03.001
- Jeoh, T., Cardona, M. J., Karuna, N., Mudinoor, A. R., and Nill, J. (2017). Mechanistic kinetic models of enzymatic cellulose hydrolysis—A review. *Biotechnol. Bioeng.* 114, 1369–1385. doi:10.1002/bit.26277
- Jeun, J.-P., Lee, B.-M., Lee, J.-Y., Kang, P.-H., and Park, J.-K. (2015). An irradiation-alkaline pretreatment of kenaf core for improving the sugar yield. *Renew. Energy* 79, 51–55. doi:10.1016/j.renene.2014.10.030
- Jin, Y., and Branke, J. (2005). Evolutionary optimization in uncertain environments—a survey. *IEEE Transactions on Evolutionary Computation* 9, 303–317. doi:10.1109/TEVC.2005.846356
- Kengpol, A., and Klaiklueng, A. (2024). Design of machine learning for limes classification based upon Thai agricultural standard no. TAS 27-2017. *Appl. Sci. Eng. Prog.* 18. doi:10.14416/j.asep.2024.01.005
- Kumar, S. R., and Chinmasamy, M. P. (2025). AI-Driven detection of tomato leaf diseases for sustainable agriculture. *Appl. Sci. Eng. Prog.* 18. doi:10.14416/j.asep.2025.06.007
- Li, M., Cao, S., Meng, X., Studer, M., Wyman, C. E., Ragauskas, A. J., et al. (2017). The effect of liquid hot water pretreatment on the chemical-structural alteration and the reduced recalcitrance in poplar. *Biotechnol. Biofuels* 10, 237. doi:10.1186/s13068-017-0926-6
- Li, X., Shi, Y., Kong, W., Wei, J., Song, W., and Wang, S. (2022). Improving enzymatic hydrolysis of lignocellulosic biomass by bio-coordinated physicochemical pretreatment—A review. *Energy Rep.* 8, 696–709. doi:10.1016/j.egyrs.2021.12.015
- Namboonlue, S., Ngowsakul, K., Nakarat, K., Kongsinkaew, C., Subjalearndee, N., Uttayopas, P., et al. (2025). Predictive reducing sugar release from lignocellulosic biomass using sequential acid pretreatment and enzymatic hydrolysis by harnessing a machine learning approach. *Comput. Struct. Biotechnol. J.* 27, 4246–4256. doi:10.1016/j.csbj.2025.09.027
- Nguyen, V. (2019). “Bayesian optimization for accelerating hyper-parameter tuning,” in 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), 302–305. doi:10.1109/AIKE.2019.00060
- Niu, Y., Joseph, N., Hemashini, T., and Leh, C. P. (2024). Valorization of lignocellulosic biomass: progress in the production of second-generation bioethanol. *Renew. Energies* 2, 27533735241284221. doi:10.1177/27533735241284221
- Niu, Y., Lee, C. K., and Leh, C. P. (2025a). Kapok pod fibre as a sustainable biofuel resource: prediction of cellulose hydrolysis using a heuristic algorithm optimized random forest. In: Y. Zhu, editor. *Proceedings of the 4th international symposium in environmental science and industrial ecology*. Singapore: Springer Nature. p. 205–214. doi:10.1007/978-981-96-1578-0\_15
- Niu, Y., Tye, Y. Y., Hemashini, T., and Leh, C. P. (2025b). Feasibility of use limited data to establish a relationship between chemical composition and the enzymatic glucose yield using machine learning. *Biomass Bioenergy* 200, 107956. doi:10.1016/j.biombioe.2025.107956
- Öhgren, K., Bura, R., Saddler, J., and Zacchi, G. (2007). Effect of hemicellulose and lignin removal on enzymatic hydrolysis of steam pretreated corn stover. *Bioresour. Technol.* 98, 2503–2510. doi:10.1016/j.biortech.2006.09.003
- Probst, P., Wright, M. N., and Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Min. Knowl. Discov.* 9, e1301. doi:10.1002/widm.1301
- Quinn, T. P., Erb, I., Gloor, G., Notredame, C., Richardson, M. F., and Crowley, T. M. (2019). A field guide for the compositional analysis of any-omics data. *Gigascience* 8, giz107. doi:10.1093/gigascience/giz107
- Rakshit, P., Konar, A., and Das, S. (2017). Noisy evolutionary optimization algorithms – a comprehensive survey. *Swarm Evol. Comput.* 33, 18–45. doi:10.1016/j.swevo.2016.09.002
- Ramaraj, R., and Unpaprom, Y. (2019). Optimization of pretreatment condition for ethanol production from *Cyperus difformis* by response surface methodology. *3 Biotech.* 9, 218. doi:10.1007/s13205-019-1754-0
- Rezaee Jordehi, A., and Jasni, J. (2015). Particle swarm optimisation for discrete optimisation problems: a review. *Artif. Intell. Rev.* 43, 243–258. doi:10.1007/s10462-012-9373-8
- Saju, L., Selvaraj, D., and Vairaperumal, T. (2025). Chapter 9 - artificial intelligence and machine intelligence: modeling and optimization of bioenergy production. In: A. K. Dubey, A. L. Srivastav, A. Kumar, U. C. Pati, F. P. García Márquez, and V. García-Díaz, editors. *Computer vision and machine intelligence for renewable energy systems*. Berlin: Elsevier. p. 163–176. doi:10.1016/B978-0-443-28947-7.00009-4
- Sharma, P., and Sharma, N. (2024). RSM approach to pre-treatment of lignocellulosic waste and a statistical methodology for optimizing bioethanol production. *Waste Manag. Bull.* 2, 49–66. doi:10.1016/j.wmb.2023.12.004
- Sluiter, A., Hames, B., Ruiz, R., Scarlata, C., Sluiter, J., Templeton, D., et al. (2008). Determination of structural carbohydrates and lignin in biomass. Golden, Colorado, U.S: National Renewable Energy Laboratory. Available online at: <https://permanent.access.gpo.gov/lps94089/42618.pdf>.
- Tabish, A. N., Irfan, M., Irshad, M., Hussain, M. A., Zeb, H., Jahangir, S., et al. (2024). Optimization of waste biomass demineralization through response surface methodology and enhancement of thermochemical and fusion properties. *Sci. Rep.* 14, 27246. doi:10.1038/s41598-024-63471-4
- Tajuddin, M., Ahmad, Z., and Ismail, H. (2016). A review of natural fibers and processing operations for the production of binderless boards. *BioRes* 11, 5600–5617. doi:10.15376/biores.11.2.Tajuddin
- Tye, Y. Y., Lee, K. T., Wan Abdullah, W. N., and Leh, C. P. (2013). Potential of *Ceiba pentandra* (L.) Gaertn. (kapok) fiber as a resource for second generation bioethanol: parametric optimization and comparative study of various pretreatments prior enzymatic saccharification for sugar production. *Bioresour. Technol.* 140, 10–14. doi:10.1016/j.biortech.2013.04.069
- Tye, Y. Y., Lee, K. T., Wan Abdullah, W. N., and Leh, C. P. (2016). Optimization of various pretreatments condition of kenaf core (*Hibiscus cannabinus*) fibre for sugar production: effect of chemical compositions of pretreated fibre on enzymatic hydrolysability. *Renew. Energy* 99, 205–215. doi:10.1016/j.renene.2016.06.040
- Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinforma.* 7, 91. doi:10.1186/1471-2105-7-91
- Wang, L., Li, S., Tian, F., and Fu, X. (2004). A noisy chaotic neural network for solving combinatorial optimization problems: stochastic chaotic simulated annealing. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 34, 2119–2125. doi:10.1109/TSMCB.2004.829778
- Xie, T., and Fan, M. (2025). Machine learning models for predicting enzymatic hydrolysis yields of lignocellulosic biomass after various pretreatments. *Industrial Crops Prod.* 235, 121644. doi:10.1016/j.indcrop.2025.121644
- Yang, B., and Wyman, C. E. (2008). Pretreatment: the key to unlocking low-cost cellulosic ethanol. *Biofuels Bioprod. Bioref.* 2, 26–40. doi:10.1002/bbb.49
- Ying Tye, Y., Peng Leh, C., Teong Lee, K., and Wan Abdullah, W. N. (2017). Non-wood lignocellulosic biomass for cellulosic ethanol production: effects of pretreatment on chemical composition in relation to total glucose yield. *J. Jpn. Inst. Energy* 96, 503–508. doi:10.3775/jie.96.503
- Yuan, Q., Liu, S., Ma, M.-G., Ji, X.-X., Choi, S.-E., and Si, C. (2021a). The kinetics studies on hydrolysis of hemicellulose. *Front. Chem.* 9, 781291. doi:10.3389/fchem.2021.781291
- Yuan, Y., Jiang, B., Chen, H., Wu, W., Wu, S., Jin, Y., et al. (2021b). Recent advances in understanding the effects of lignin structural characteristics on enzymatic hydrolysis. *Biotechnol. Biofuels* 14, 205. doi:10.1186/s13068-021-02054-1