Jurnal Teknik Industri, Vol. 28, No. 2, December 2026
ISSN 1411-2485 print / ISSN 2087-7439 online
DOI: https://doi.org/10.9744/jti.28.2.152-163

Bayesian-Optimized Gaussian Process Regression for Robust Prediction of Biodiesel–Alcohol Blend Density under Thermophysical Group-Based Validation

Ahmad Setyo Putra, Areta Triani Putri, Rudi Nurdiansyah*, Andoko
Mechanical and Industrial Engineering Department, Universitas Negeri Malang,
Jl. Semarang No. 5, Malang 65145, Indonesia
Email: rudi.nurdiansyah.ft@um.ac.id*
*Corresponding author

Abstract: Accurate density prediction is important for biodiesel–alcohol blend development; however, random train-test splitting may overestimate the performance in structured thermophysical datasets. This study evaluated whether Bayesian-optimized Gaussian Process Regression (BO-GPR) can be generalized to unseen thermophysical groups. A dataset of 1,856 observations was assessed using random row-wise split, group temperature, group pressure, and group alcohol composition validation. Seven regression models were compared using 10 repeated outer splits, validation-specific inner cross-validation, and Bayesian hyperparameter optimization. Model differences were examined using paired Wilcoxon signed-rank tests with Holm correction, and GPR uncertainty was assessed using 95% prediction-interval coverage. The BO-GPR achieved median R^2 values of 0.99997, 0.99864, 0.99999, and 0.99582 across the four validation schemes, with corresponding median RMSE values of 0.00024, 0.00143, 0.00012, and 0.00246 g/cm^3, respectively. The GPR showedPR showed a significantly lower RMSE than all the benchmark models under group-based validation, whereas the difference from CatBoost was not significant under random splitting. The prediction interval coverage remained below the nominal 95% level, indicating that uncertainty calibration requires further improvement. These findings support the use of BO-GPR as a predictive surrogate within the evaluated domain, although independent experimental validation is required for broader applications.

Keywords: Bayesian optimization, biodiesel–alcohol blends, Gaussian process regression, machine learning.

Introduction

The development of biodiesel and biodiesel–alcohol blends has become increasingly important in renewable energy research because of the need for alternative fuels that are more sustainable, compatible with diesel engines, and capable of providing stable combustion characteristics compared to conventional fossil fuels. In the formulation of these fuels, density is a key thermophysical property that affects fuel injection, atomization, spray penetration, spray angle, droplet size, combustion efficiency, storage stability, and fuel quality assessment [1], [2]. Density variations caused by changes in temperature, pressure, alcohol type, and blend composition can influence the air–fuel mixing quality and combustion performance [3]. Therefore, accurate density prediction is essential in the development of biodiesel–alcohol blends, particularly because experimental measurements can be costly, time-consuming, and limited to specific operating ranges [4], [5].

Physically, the density of biodiesel–alcohol blends are governed by the simultaneous effects of temperature, pressure, alcohol type, and composition. Density generally decreases as temperature increases owing to thermal expansion, whereas increasing pressure increases density through fluid compressibility effects [6], [7]. The addition of alcohols such as 1-butanol and 2-butanol also tends to reduce the blend density because alcohols have lower densities than biodiesel, although the magnitude of this effect depends on the molecular structure and blend fraction [8], [9], [10]. Empirical and semi-empirical models, including Tait, Tammann–Tait, PC-SAFT, Wilson, NRTL, mixing rules, and two-dimensional regression, have been used to correlate the density and other thermophysical properties of biodiesel [11], [12], [13]. However, these approaches may have limited validity ranges and may be less flexible in representing nonlinear relationships within complex temperature–pressure–composition systems.

Machine learning has emerged as a data-driven approach capable of predicting biodiesel and thermophysical properties with high accuracy. Models such as random forest, XGBoost, CatBoost, artificial neural network,

Submitted: July 24, 2026; Revised: Sept 30, 2026; Accepted: Oct 07, 2026

support vector machine, adaptive neuro-fuzzy inference system, and Gaussian process regression have been applied to predict the density, viscosity, cetane number, calorific value, cold flow properties, thermal conductivity, and other fuel properties [14], [15], [16], [17]. In the context of biodiesel–alcohol density prediction, previous studies have shown that machine learning models, such as Gaussian process modeling, ANFIS, and decision trees, can produce low prediction errors, including very small MAPE values for specific datasets [18]. In addition, the integration of machine learning with physical models or equations of state has been explored to improve the prediction of thermophysical properties and address the limitations associated with experimental data [13], [19], [20], [21]. However, a methodological gap remains in the evaluation of model performance when thermophysical datasets contain structured combinations of temperature, pressure, composition and alcohol type. Random splitting may place similar thermophysical conditions in both the training and test sets, causing overly optimistic estimates based on interpolation rather than unseen-condition generalization [22]. Therefore, group-based validation based on temperature, pressure, and alcohol composition combinations is required to assess model generalization under conditions that are structurally separated from the training data.

Based on this gap, this study developed and evaluated a Bayesian-optimized Gaussian process regression (GPR) framework for the density prediction of biodiesel–alcohol blends under random row-wise and thermophysical group-based validation. Gaussian process regression was selected because of its probabilistic formulation, ability to model nonlinear relationships, and capability to provide predictive uncertainty, whereas Bayesian optimization was employed to systematically optimize model hyperparameters within predefined search spaces [23], [24], [25], [26]. The proposed model was compared with linear regression, ridge regression, decision tree, random forest, XGBoost, and CatBoost using random row-wise validation and group-based validation based on the temperature, pressure, and alcohol composition. The main contribution of this study is the integration of Bayesian-optimized GPR with thermophysical group-based validation to produce a biodiesel–alcohol density prediction model that is not only accurate under random split evaluation but also robust in predicting unknown thermophysical conditions.

Methods

Research Framework and Dataset

This study employed a comparative quantitative modeling framework to predict the density of biodiesel–alcohol blends using experimental thermophysical data compiled from three previously published studies on waste cooking oil biodiesel blended with 1-butanol and 2-butanol under various temperature, pressure, and composition conditions [27], [28], [29]. The dataset contained two alcohol types, 12 composition levels, 23 pressure levels, and six temperature levels for the model training. The overall research framework is shown in Figure 1 and comprises data preparation, thermophysical grouping, repeated validation, model development, Bayesian hyperparameter optimization, performance evaluation, statistical comparison, and GPR uncertainty analyses.

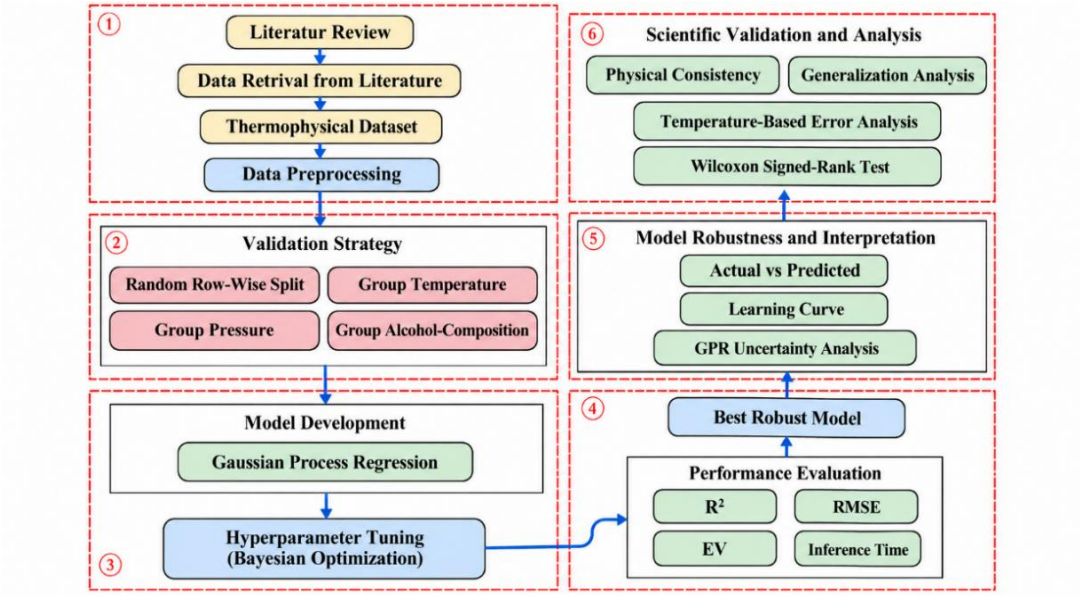


Figure 1. Research framework for Bayesian-optimized Gaussian process regression

Table 1. Thermophysical characteristics of the integrated dataset

Variable	Definition	Minimum	Maximum	Unit
Composition x	Mole fraction of WCOB	0	1	mol fraction
Pressure p	Applied pressure	0.1	140	MPa
Temperature T	Measurement temperature	298.15	393.15	K
Density ρ	Measured blend density	0.7048	0.9427	g/cm³

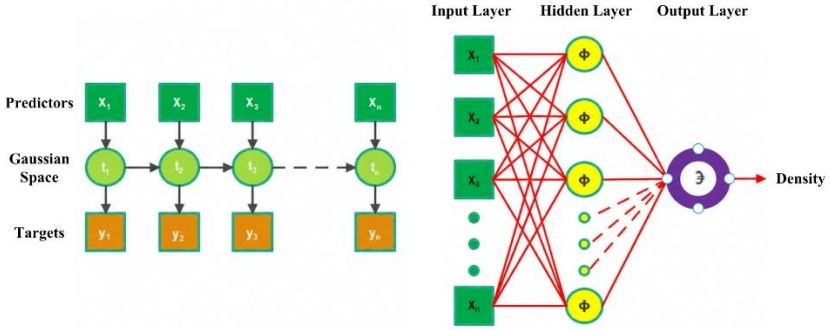


Figure 2. Gaussian process regression model

The selection of these variables was based on the thermophysical relationship in which biodiesel–alcohol blend density is influenced by composition, pressure, temperature, and alcohol type. Previous studies have shown that temperature and alcohol composition generally decrease density, whereas pressure increases density owing to fluid compressibility effects [18]. In the compiled experimental data, the composition variable x represents the mole fraction of waste cooking oil biodiesel (WCOB), with $x = 0$ representing the alcohol-rich endpoint and $x = 1$ representing pure WCOB. This definition is consistent with the experimental studies used to construct our dataset. To distinguish interpolation from generalization to unseen thermophysical conditions, four validation schemes were used: random row-wise split, group temperature, group pressure, and group alcohol–composition. Each scheme was evaluated using 10 repeated outer splits with different random seeds, and hyperparameter optimization was performed only on the corresponding training data using validation-specific inner cross-validation.

$$x_i = [x_i . p_i . T_i . a_i] \tag{1}$$

$$y_i = \rho_i \tag{2}$$

where $x_i$denotes WCOB mole fraction, p_i denotes pressure in MPa, T_i denotes temperature in K, a_i denotes alcohol type, and ρ_i denotes blend density in g/cm³. Composition, pressure, and temperature were treated as numerical variables, whereas alcohol type was treated as a categorical variable and encoded using one-hot encoding. For linear regression, ridge regression, and GPR, the numerical features were standardized using StandardScaler, whereas tree-based models used the original numerical scale. All preprocessing operations were implemented within the corresponding model pipelines and fitted using the training data only.

Gaussian Process Regression (GPR)

Gaussian process regression (GPR) was used as the primary probabilistic regression model because it can represent nonlinear relationships between thermophysical variables and density while providing both predictive means and predictive uncertainty. Unlike parametric models that assume a predefined functional form, GPR defines a distribution over latent regression functions through a covariance structure. This formulation is suitable for continuous thermophysical data, in which density depends on the combined effects of temperature, pressure, composition, and alcohol type. Similar probabilistic machine learning approaches have been applied to biodiesel property prediction using experimental datasets [30], [31]. Figure 2 illustrates the conceptual structure of the GPR model used in this study.

In this study, the relationship between the input vector (x) and density target (y) was modeled as follows:

$$\mathrm{y} = \mathrm{f}(x) + \varepsilon \tag{3}$$

where ε represents the observation noise. The latent function is assumed to follow a Gaussian process:

$$\mathrm{f}(x) \sim GP\big(m(x), k(x, x')\big) \tag{4}$$

Table 2. GPR kernel configuration

Kernel type	Kernel structure	Description
RBF	RBF	Smooth nonlinear covariance function
Matern 1.5	Matern, ν = 1.5	Less smooth covariance function
Matern 2.5	Matern, ν = 2.5	Moderately smooth covariance function
RBF + WhiteKernel	RBF + WhiteKernel	Smooth covariance function with noise component

where $m(x)$ represents the mean function and $k(x, x')$ denotes the covariance function or kernel. In the implementation, normalize_y=True was used so that the density target was normalized internally during GPR fitting. Given the training data (X) and target (y), the predictive distribution for a test point x_* was obtained from the predictive mean and variance:

$$\mu_* = k_*^T (K + \alpha I)^{-1} y \tag{5}$$

$$\sigma_*^2 = \mathrm{k}(x_*, x_*) - k_*^T (K + \alpha I)^{-1} k_* \tag{6}$$

where K is the covariance matrix of the training data, k_* is the covariance vector between the test point and the training data, α is the diagonal noise/regularization parameter, and I is the identity matrix. α is added to the diagonal of the covariance matrix to improve numerical stability and can also represent additional observation noise. Therefore, α is interpreted primarily as a statistical regularization/noise parameter rather than as a direct physical property of the biodiesel–alcohol blend. In this study, α was optimized within the range (10^{-8}) to (10^{-2}) using Bayesian optimization for each validation scheme. Because the target was normalized during GPR fitting, differences in optimized α between validation schemes are interpreted in terms of the required regularization of the fitted model rather than as direct differences in physical measurement noise. The predictive variance was subsequently used to construct empirical 95% prediction intervals for uncertainty analysis.

The GPR kernel configurations are presented in Table 2. Four kernel structures were evaluated: RBF, Matern 1.5, Matern 2.5, and RBF with WhiteKernel, which represent smooth, different degrees of covariance smoothness, and explicit noise effects, respectively. These kernels were selected to provide a focused and computationally tractable search space for continuous thermophysical data, whereas other kernel families were excluded to limit the optimization scope.

The RBF kernel is formulated as

$$k_{RBF}(x, x') = \sigma_f^2 \exp\left(-\frac{|x - x'|^2}{2l^2}\right) \tag{7}$$

where σ_f^2 is the signal variance and (l) is the length scale. Meanwhile, the Matern kernel is expressed as:

$$k_{Matern}(x, x') = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu} r}{l}\right) \tag{8}$$

where $r = |x - x'|$, (ν) is the kernel smoothness parameter, Γ(ν) is the gamma function, and (K_ν) is the modified Bessel function. The value ν = 1.5 was used for a less smooth kernel, while ν = 2.5 was used for a covariance function with moderate smoothness. In the RBF + WhiteKernel configuration, the covariance structure was supplemented with an explicit noise component, as follows:

$$\mathrm{k}(x, x') = k_{RBF}(x, x') + \sigma_n^2 \delta_{ij} \tag{9}$$

where σ_n^2 is the noise variance and δ_{ij} is the Kronecker delta. This structure was used to evaluate whether adding an explicit noise component could improve model robustness against variations in experimental data.

Bayesian Optimization

Bayesian optimization (BO) was implemented using BayesSearchCV with 100 iterations, RMSE-based scoring, model-specific search spaces, a different random seed for each repeated outer split, and parallel computation using all available CPU cores (n_jobs = −1). The optimizer used its implementation defaults for the initial evaluations, surrogate model, and acquisition strategies. For random row-wise validation, 5-fold shuffled KFold was used, whereas GroupKFold was applied for group-based validation. Hyperparameter optimization was

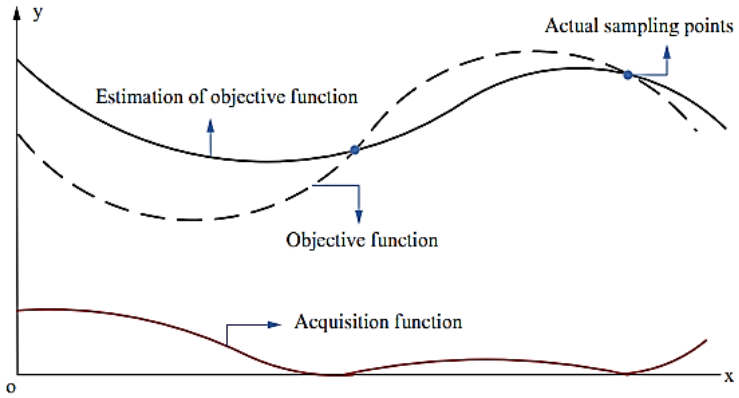


Figure 3. Schematic of Bayesian optimization

Table 3. Hyperparameter search space of Bayesian-optimized GPR

Hyperparameter	Search space	Description
alpha	$10^{-8} - 10^{-2}$, Log uniform	Noise regularization parameter for numerical stability and model smoothness
kernel_type	RBF, Matern 1.5, Matern 2.5, RBF + WhiteKernel	Covariance structure for nonlinear density modeling

Table 4. Machine learning models used for comparative density prediction

Category	Model	Role	Optimization
Baseline regression	Linear regression (LR)	Untuned linear reference	None
Regularized regression	Ridge regression (RR)	Regularized linear benchmark	Bayesian optimization
Tree-based model	Decision tree (DT)	Nonlinear benchmark	Bayesian optimization
Ensemble model	Random forest (RF)	Bagging-based benchmark	Bayesian optimization
Boosting model	XGBoost (XGB)	Gradient boosting benchmark	Bayesian optimization
Boosting model	CatBoost (CB)	Ordered boosting benchmark	Bayesian optimization
Probabilistic model	Gaussian Process Regression (GPR)	Proposed probabilistic model	Bayesian optimization

restricted to the corresponding outer training data, followed by refitting and evaluation of the held-out outer test set. Figure 3 illustrates the general principle of BO for estimating the objective function and selecting the next sampling point using an acquisition function. In this study, BO was applied to optimize the GPR hyperparameters, particularly (α) and kernel type.

In this study, the hyperparameter search space is shown in Table 3, which included α from (10^{-8}) to (10^{-2}) using a log-uniform distribution and four candidate kernel structures: RBF, Matern 1.5, Matern 2.5, and RBF + WhiteKernel. The same Bayesian optimization framework was applied to Ridge Regression, Decision Tree, Random Forest, XGBoost, and CatBoost using model-specific search spaces, while Linear Regression was retained as an untuned reference baseline. In biodiesel machine learning studies, Bayesian-based optimization approaches have also been used to improve model accuracy and reduce the need for manual hyperparameter search [32], [33].

In general, the optimization process aims to find the best hyperparameter configuration (θ) which minimizes the cross-validation error. The objective function is formulated as follows:

$$\theta^* = \arg\min_{\theta \in \Theta} RMSE_{CV}(\theta) \tag{10}$$

where $(\theta = \alpha, kernel_type)$, (Θ) is the hyperparameter search space, and $(\mathcal{L}_{CV})$ is the loss value obtained from cross-validation. The implementation used neg_root_mean_squared_error; therefore, maximizing the negative RMSE is equivalent to minimizing cross-validation RMSE. A budget of 100 iterations was selected as a practical balance between search coverage and computational cost across repeated validation experiments and multiple benchmark models.

Benchmark Models

To evaluate the proposed GPR framework, it was compared with linear regression, ridge regression, decision tree, random forest, XGBoost, and CatBoost, as shown in Table 4. Linear regression was retained as an untuned reference baseline, whereas the other benchmark models were optimized using Bayesian optimization with

Table 5. Validation schemes used to evaluate interpolation and generalization performance

Scheme	Outer strategy	Inner strategy	Purpose
Random row-wise	Random 80:20 row-wise split	5-fold shuffled KFold	Interpolation
Group temperature	GroupShuffleSplit, 20% test-group proportion	GroupKFold	Unseen temperature levels
Group pressure	GroupShuffleSplit, 20% test-group proportion	GroupKFold	Unseen pressure levels
Group alcohol–composition	GroupShuffleSplit, 20% test-group proportion	GroupKFold	Unseen alcohol–composition groups

model-specific search spaces. Linear models represent conventional regression relationships, whereas tree-based and boosting models provide nonlinear benchmarks commonly used for tabular prediction and biodiesel property modeling [18], [34]. The comparison was performed under both random row-wise interpolation and repeated group-based validation of unseen temperature, pressure, and alcohol–composition conditions.

Validation Strategy and Evaluation Metrics

The validation strategy comprised random row-wise split, group temperature, group pressure, and group alcohol–composition. Random splitting used train_test_split with 20% test data to assess interpolation, whereas group-based validation used GroupShuffleSplit with a 20% test-group proportion: temperature values rounded to two decimals formed six groups, pressure values rounded to four decimals formed 23 groups, and each alcohol type combined with the WCOB mole fraction rounded to four decimals formed 14 groups. Group overlap between the outer training and test sets was checked for every split. For example, an observation at 310.15 K was assigned to the 310.15-K temperature group, an observation at 20.0000 MPa to the 20.0000-MPa pressure group, and an observation with 1-butanol and WCOB mole fraction 0.4999 to the alcohol–composition group 1-butanol_x=0.4999. Each scheme was repeated 10 times using seeds 42–51; 5-fold shuffled KFold was used for random validation and GroupKFold for group-based validation, with hyperparameter optimization restricted to the outer training data. Model performance was evaluated using R^2, RMSE, EV, and inference time, with RMSE serving as the primary error metric for Bayesian optimization and Wilcoxon signed-rank comparison with Holm correction.

The Model performance was evaluated using the coefficient of determination R^2, RMSE, explained variance (EV), inference time, and tuning time. The evaluation metrics were formulated as follows:

$$\mathrm{EV} = 1 - \frac{Var(y-\hat{y})}{Var(y)} \tag{11}$$

$$t_{inf} = (t_{end} - t_{start}) \times 1000 \tag{12}$$

Equation (11) defines explained variance (EV), which quantifies the proportion of variance in the observed target values accounted for by the model predictions. Equation (12) calculates inference time in milliseconds based on the elapsed time between the start and end of the prediction process. To assess whether differences in predictive error between GPR and the benchmark models were statistically significant, paired Wilcoxon signed-rank tests were performed using RMSE values from the same 10 repeated outer splits. The Holm–Bonferroni correction was applied within each validation scheme to account for multiple pairwise comparisons, with a significance level of 0.05.

Results and Discussions

Bayesian Optimization Results

The Bayesian optimization results for the four validation schemes are presented in Table 6. The selected kernel varied across validation schemes and repeated runs, with RBF + WhiteKernel most frequently selected for Random Row-Wise Split (50%), Matern 1.5 for Group Temperature (40%), Matern 2.5 for Group Pressure (40%), and RBF for Group Alcohol composition (50%). This variation indicates that the preferred covariance structure depends on the validation scenario rather than a single kernel being consistently optimal for all the scenarios.

The optimized α values also depend on the validation scheme. Random Row-Wise Split generally selected values around 10^{-6}, whereas Group Temperature and Group Alcohol–Composition frequently selected the upper search boundary of 10^{-2}, indicating a stronger preference for regularization under unseen-condition validation.

Table 6. Repeated Bayesian optimization results for GPR across validation schemes

Validation scheme	Kernel selection	Alpha behavior
Random row-wise	RBF + WhiteKernel (50%)	Generally in the order of 10^{-6}
Group temperature	Matern 1.5 (40%)	10^{-2} selected in 8/10 runs
Group pressure	Matern 2.5 (40%)	Generally low, variation across runs
Group alcohol–composition	RBF (50%)	10^{-2} selected in 8/10 runs

Table 7. Median R^2 and interquartile range across validation schemes

Model	Random R^2	Group temp. R^2	Group pressure R^2	Group comp. R^2
GPR	0.99997 [0.99995–0.99999]	0.99864 [0.99669–0.99948]	0.99999 [0.99997–0.99999]	0.99582 [0.99155–0.99895]
LR	0.98306 [0.98194–0.98353]	0.98172 [0.97900–0.98616]	0.97960 [0.97844–0.98191]	0.98196 [0.97471–0.98518]
RR	0.98306 [0.98193–0.98353]	0.98158 [0.97817–0.98565]	0.98001 [0.97849–0.98203]	0.98196 [0.97396–0.98544]
DT	0.98740 [0.98669–0.98845]	0.85853 [0.83147–0.86665]	0.98912 [0.98764–0.99010]	0.94710 [0.93699–0.96262]
RF	0.99599 [0.99569–0.99620]	0.86444 [0.84429–0.87791]	0.99042 [0.99000–0.99151]	0.94720 [0.93328–0.96289]
XGB	0.99936 [0.99934–0.99946]	0.85785 [0.63850–0.86492]	0.98379 [0.98193–0.98898]	0.95008 [0.91208–0.95953]
CB	0.99994 [0.99994–0.99995]	0.86055 [0.84213–0.86602]	0.98932 [0.98860–0.99197]	0.98375 [0.96549–0.99013]

Table 8. Median RMSE and interquartile range across validation schemes

Model	Random RMSE	Group temp. RMSE	Group pressure RMSE	Group comp. RMSE
GPR	0.00024 [0.00015–0.00030]	0.00143 [0.00084–0.00192]	0.00012 [0.00009–0.00025]	0.00246 [0.00116–0.00380]
LR	0.00543 [0.00515–0.00563]	0.00453 [0.00440–0.00583]	0.00577 [0.00507–0.00621]	0.00517 [0.00477–0.00604]
RR	0.00543 [0.00515–0.00563]	0.00453 [0.00448–0.00595]	0.00575 [0.00503–0.00624]	0.00511 [0.00467–0.00613]
DT	0.00469 [0.00441–0.00473]	0.01377 [0.01362–0.01399]	0.00418 [0.00405–0.00430]	0.00851 [0.00749–0.00929]
RF	0.00264 [0.00255–0.00273]	0.01328 [0.01306–0.01382]	0.00372 [0.00351–0.00421]	0.00830 [0.00697–0.00959]
XGB	0.00106 [0.00094–0.00108]	0.01383 [0.01372–0.02042]	0.00480 [0.00426–0.00559]	0.00908 [0.00747–0.01182]
CB	0.00031 [0.00028–0.00033]	0.01368 [0.01341–0.01382]	0.00391 [0.00353–0.00445]	0.00493 [0.00370–0.00691]

The previously reported values of approximately $10^0 - 10^4$ were inconsistent with the declared search space and were therefore corrected; α is a model regularization/noise parameter rather than a direct measure of physical experimental noise. Because the upper bound was frequently selected, these values represent the preferred regularization within the tested range rather than an optimum unconstrained value.

Comparative Model Performance

The results in Tables 7 and 8 and Figures 4 and 5 show that GPR achieved the highest predictive performance across the four validation schemes. The median R^2 The values were 0.99997, 0.99864, 0.99999, and 0.99582 for Random Row-Wise Split, Group Temperature, Group Pressure, and Group Alcohol composition, respectively, with corresponding median RMSE values of 0.00024, 0.00143, 0.00012, and 0.00246 g/cm³. The largest performance differences occurred under Group Temperature, where tree-based and boosting models showed substantially lower median R^2, particularly XGBoost (0.85785), whereas Linear Regression and Ridge remained at approximately 0.982. These results indicate that group-based validation reveals differences in model generalization that are not apparent under random row-wise splitting of the dataset.

The paired Wilcoxon signed-rank results in Table 9 provide statistical support for these differences. After Holm correction, GPR showed significantly lower RMSE than all benchmark models under Group Temperature, Group Pressure, and Group Alcohol–Composition ($P_{adj} = 0.0195$). Under Random Row-Wise Split, GPR also showed significantly lower RMSE than Linear Regression, Ridge, Decision Tree, Random Forest, and XGBoost ($P_{adj} = 0.0117$), whereas its difference from CatBoost was not statistically significant ($P_{adj} = 0.1055$). Thus,

Table 9. Paired Wilcoxon signed-rank test comparing GPR and benchmark models based on RMSE

Validation scheme	LR	Ridge	DT	RF	XGB	CB
Random row-wise	0.0117*	0.0117*	0.0117*	0.0117*	0.0117*	0.1055
Group temperature	0.0117*	0.0117*	0.0117*	0.0117*	0.0117*	0.0117*
Group pressure	0.0117*	0.0117*	0.0117*	0.0117*	0.0117*	0.0117*
Group alcohol–composition	0.0117*	0.0117*	0.0117*	0.0117*	0.0117*	0.0195*

Note: Values are Holm-adjusted p-values from paired Wilcoxon signed-rank tests using 10 paired repeats. An asterisk (*) indicates statistical significance at the 0.05 level $(p_{adj} < 0.05)$; unmarked values are not statistically significant.

the statistical evidence for the GPR advantage is strongest under group-based validation, where models are evaluated on thermophysical groups excluded from training.

Figures 4 and 5 illustrate the median R^2 and median RMSE, respectively, with their corresponding interquartile ranges. GPR maintained median R^2 values above 0.99 across all four validation schemes and generally showed narrower interquartile ranges than the tree-based and boosting models under group-based validation. Linear Regression and Ridge maintained median R^2 values around 0.982 under Group Temperature, suggesting that a systematic component of the temperature–density relationship can be captured by simpler regression models. In contrast, the lower median R^2 and wider interquartile ranges of the tree-based and boosting models indicate greater sensitivity to changes in the temperature conditions represented in the test data.

The observed model behavior was consistent with the thermodynamic characteristics of the biodiesel–alcohol blends. The density generally decreases with increasing temperature because of thermal expansion, whereas increasing pressure increases the density through compressibility effects. The low median RMSE under Group Pressure (0.00012 g/cm³) indicates that GPR predicted density accurately for the held-out pressure groups, whereas the higher median RMSE under Group Temperature (0.00143 g/cm³) reflects the greater difficulty in generalizing across unseen temperature levels. Overall, GPR maintained median R^2 values above 0.99 and the lowest median RMSE across the four validation schemes, demonstrating the importance of evaluating model performance with group-based validation rather than relying solely on random row-wise splitting of the data.

Robustness and Error Analysis

Figure 6 presents the actual versus predicted density for BO-GPR under a representative group temperature split, with most observations concentrated around the diagonal line. This behavior is consistent with the repeated validation results, where GPR achieved mean R^2 The values were 0.99996, 0.99538, 0.99974, and 0.99282 for random row-wise split, group temperature, group pressure, and group alcohol–composition, respectively, with corresponding RMSE values of 0.000236, 0.001880, 0.000332, and 0.002717 g/cm³. Thus, although unseen temperature and alcohol composition conditions remain more challenging, the repeated results show that prediction accuracy is maintained across the evaluated validation schemes.

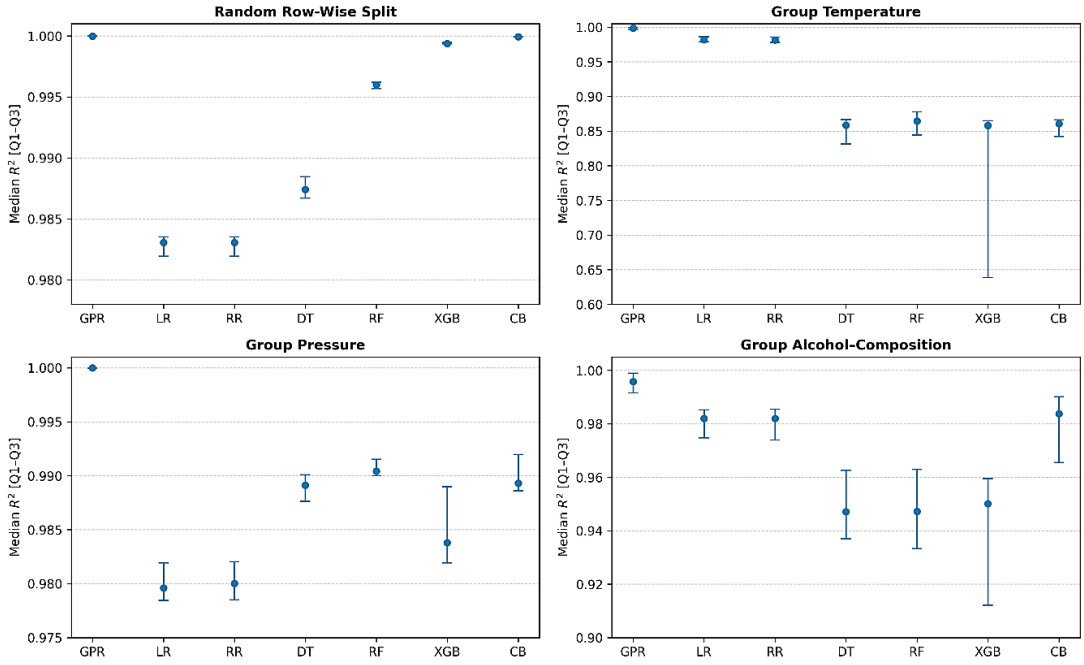


Figure 4. Median R^2 and interquartile range (Q1–Q3) across four validation schemes

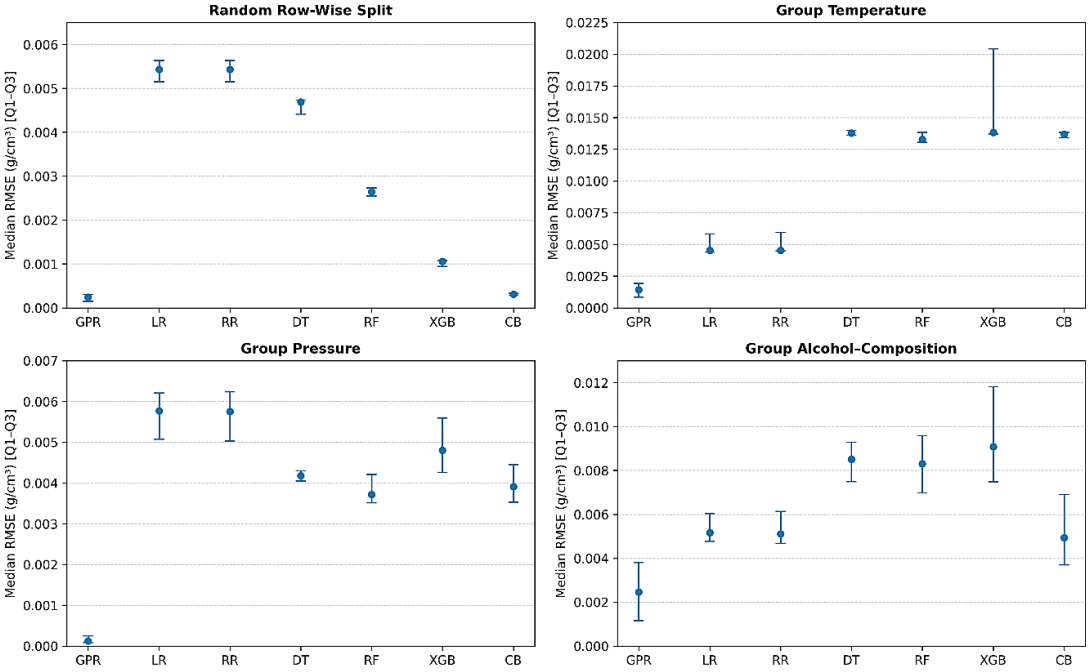


Figure 5. Median RMSE and interquartile range (Q1–Q3) across the four validation schemes

The error pattern was consistent with the thermophysical behavior of the biodiesel–alcohol blends. Temperature affects density through thermal expansion; therefore, excluding temperature levels from the model training requires the GPR model to infer the density response from neighboring thermophysical patterns. Figure 7 shows the RMSE at the two temperature levels in the representative group temperature split. The RMSE was 0.00617 g/cm³ at 298.15 K and 0.00055 g/cm³ at 313.15 K, indicating that the prediction error differed between the two held-out temperature levels. This behavior is also consistent with conventional thermodynamic density correlations and equations of state, and mixture-based approaches reported for biodiesel and biofuel systems [27], [28]. The present framework achieved a mean R^2 of 99.70% across the four validation schemes; however, the results are not directly comparable because the datasets and validation protocols differ.

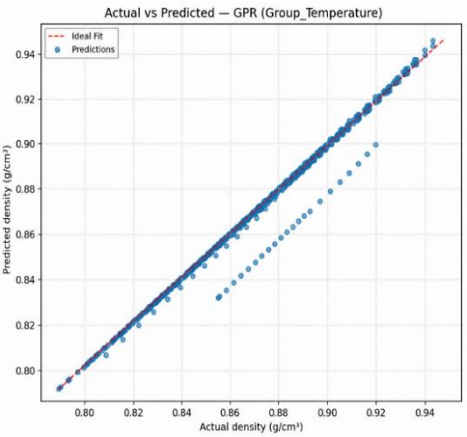


Figure 6. Actual versus predicted density for BO-GPR under a representative group temperature split

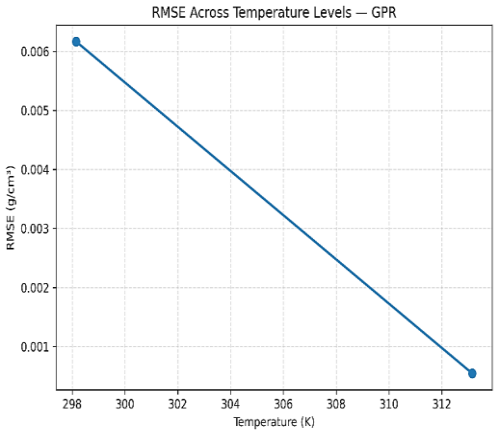


Figure 7. RMSE at the two temperature levels for BO-GPR under a representative group temperature split

Table 10. BO-GPR predictive performance and computational cost across validation schemes

Validation scheme	R^2 mean ± sd	RMSE mean ±sd (g/cm³)	EV mean ± sd	Inference time (ms) mean ± sd	Tuning time (s) mean ± sd
Random row-wise split	0.99996 ± 0.00003	0.000236 ± 0.000090	0.99996 ± 0.00003	9.42 ± 0.13	1454.82 ± 109.93
Group temperature	0.99538 ± 0.00684	0.001880 ± 0.001378	0.99577 ± 0.00650	11.53 ± 0.16	592.26 ± 116.87
Group pressure	0.99974 ± 0.00078	0.000332 ± 0.000576	0.99974 ± 0.00078	9.45 ± 0.36	1664.61 ± 289.81
Group alcohol–composition	0.99282 ± 0.00848	0.002717 ± 0.001786	0.99391 ± 0.00768	9.62 ± 0.25	1057.26 ± 272.14

The GPR predictive intervals provided empirical coverage below the nominal 95% level across the evaluated validation schemes, indicating that the uncertainty estimates were not fully calibrated. The prediction interval width varied across the validation schemes, reflecting differences in predictive uncertainty under random and unseen condition validations. Therefore, although GPR provides probabilistic predictions, the resulting intervals should be interpreted as uncalibrated empirical uncertainty estimates within the present experimental domain

Table 10 summarizes the repeated BO-GPR performance and the computational costs. The inference time ranged from 9.42 to 11.53 ms across the four validation schemes, whereas Bayesian optimization required substantially more computational time, ranging from 592.26 to 1664.61 s. The experiments were conducted using Google Colab connected to a Windows 11 local runtime equipped with a 24-core/32-thread CPU, 31.79 GB RAM, NVIDIA GeForce RTX 5060 Ti GPU with 16 GB memory, and Python 3.13.14. These results indicate that once training and hyperparameter optimization are completed, density prediction requires only a few milliseconds for the complete held-out test set of each outer split, whereas the principal computational cost occurs during model development.

Conclusions

In this study, a Bayesian-optimized Gaussian Process Regression (BO-GPR) framework was developed for biodiesel–alcohol blend density prediction using thermophysical group-based validation. The integrated dataset comprised 1,856 observations with composition, pressure, temperature, alcohol type as input features, and density as the prediction target. Across the four validation schemes, the GPR achieved median R^2 values of 0.99997, 0.99864, 0.99999, and 0.99582 under random row-wise split, group temperature, group pressure, and group alcohol–composition, respectively. The corresponding median RMSE values were 0.00024, 0.00143, 0.00012, and 0.00246 g/cm^3. Paired Wilcoxon signed-rank tests with Holm correction confirmed significantly lower RMSE for GPR than all benchmark models under group-based validation, whereas its difference from CatBoost under random row-wise split was not statistically significant. These findings demonstrate the importance of evaluating prediction models on unseen thermophysical groups rather than relying solely on random splitting. The millisecond-level inference time supports the use of BO-GPR as a computational surrogate within the investigated data domain. However, the empirical coverage of the nominal 95% prediction intervals remained below the target level, indicating that uncertainty calibration requires further improvement. Future work should evaluate the framework using independent experimental data and calibrated uncertainty methods to assess its transferability beyond the present thermophysical domain.

Acknowledgment

This research was supported by the Institute for Research and Community Service (LPPM), State University of Malang, through the 2025 Student Innovation Grant under Contract Number 24.2.875/UN32.14.1/LT/2025. The authors gratefully acknowledge this support, which contributed to the successful completion of the study.

AI Declaration

The authors declare that AI-assisted tools (ChatGPT) were used only for minor language editing to improve clarity and readability. These tools were not used for the research design, data analysis, model development, mathematical formulation, or interpretation of the results. All scientific content, methodology, experiments, and conclusions in this manuscript were developed entirely by the authors

References

[1] H. Chen, Y. Zhang, X. Luan, M. Zhang, and G. Tao, "Experimental analysis and modeling of ethanol-biodiesel-diesel blends injection behavior," *Fluid Dynamics and Materials Processing*, vol. 21, no. 7, pp. 1753–1770, 2025, doi: https://doi.org/10.32604/fdmp.2025.066494.

[2] V. Shanthan, J. Suryawanshi, R. Tarodiya, A. Loyte, Y. Devarajan, and N. Kaliappan, "Numerical analysis of spray characterization of blends of hydrous ethanol with diesel and biodiesel," *Scientific Reports*, vol. 14, no. 1, 2024, doi: https://doi.org/10.1038/s41598-024-56444-0.

[3] F. Zheng and H. M. Cho, "Exploring the effects of synergistic combustion of alcohols and biodiesel on combustion performance and emissions of diesel engines: A review," *Energies*, vol. 17, no. 24, p. 6274, Dec. 2024, doi: https://doi.org/10.3390/en17246274.

[4] R. Aitbelale, I. Abala, F. E. M'hamdi Alaoui, A. S. Eddine, N. M. Rujas, and F. Aguilar, "Characterization and determination of thermodynamic properties of waste cooking oil biodiesel: Experimental, correlation and modeling density over a wide temperature range up to 393.15 and pressure up to 140 MPa," *Fluid Phase Equilibria*, vol. 497, pp. 87–96, 2019, doi: https://doi.org/10.1016/j.fluid.2019.06.003.

[5] C. Bouchot *et al.*, "High pressure density for dodecane + biodiesel produced from residuals from cooking oil and beef tallow: Experiments and correlation with a proposed global (p, T, x) Tammann-Tait base model," *Journal of Chemical Thermodynamics*, vol. 212, 2026, doi: https://doi.org/10.1016/j.jct.2025.107561.

[6] A. A. A. Alves *et al.*, "Investigating the thermophysical properties of the 1-Butanol + biodiesel system: The impact of pressure on volumetric characteristics," *Fuel*, vol. 371, 2024, doi: https://doi.org/10.1016/j.fuel.2024.132076.

[7] A. G. M. Ferreira, J. B. Santos, J. Baptista, S. Khalighi, R. M. M. Brito, and P. F. Cruz, "Effect of Isobutanol addition on the biodiesel density," *Journal of Chemical and Engineering Data*, vol. 66, no. 12, pp. 4542–4562, 2021, doi: https://doi.org/10.1021/acs.jced.1c00588.

[8] F. Yang, X. Wang, H. Tan, S. He, and Z. Liu, "Experimental investigations on the thermophysical properties of methyl myristate in alcoholic solutions," *Fuel*, vol. 215, pp. 187–195, 2018, doi: https://doi.org/10.1016/j.fuel.2017.10.122.

[9] K. Yadav, S. Almansour, L. M. Alkwai, A. Yadav, and F. Ranjbar, "Reliable models for density in waste cooking oil biodiesel-alcohol mixtures using advanced learning techniques," *Industrial Crops and Products*, vol. 234, p. 121602, Oct. 2025, doi: https://doi.org/10.1016/j.indcrop.2025.121602.

[10] S. Malik, P. J. Darolia, S. K. Garg, and V. K. Sharma, "Densities and excess molar volumes of mixtures containing diesel, biodiesel and alkanols at temperatures from 288.15 to 313.15 K," *Chinese Journal of Chemical Engineering*, vol. 34, pp. 198–207, Jun. 2021, doi: https://doi.org/10.1016/j.cjche.2020.09.065.

[11] P. M. Florido, P. C. F. Visioli, C. N. Pinto, and C. B. Gonçalves, "Study of FAME model systems: Database and evaluation of predicting models for biodiesel physical properties," *Renewable Energy*, vol. 151, pp. 837–845, 2020, doi: https://doi.org/10.1016/j.renene.2019.11.083.

[12] M. Gülüm and A. Bilgin, "Two-dimensional surface models to predict the density of biodiesel-diesel-alcohol ternary blends," *Energy Sources, Part A: Recovery, Utilization and Environmental Effects*, vol. 43, no. 5, pp. 517–587, 2021, doi: https://doi.org/10.1080/15567036.2019.1630034.

[13] J. Wang, Q. An, D. Zhou, K. Gao, and S. Bi, "Prediction methods for high pressure density of biofuels: A hybrid data-driven ML-PC-SAFT equation of state," *Fuel*, vol. 418, 2026, doi: https://doi.org/10.1016/j.fuel.2026.138721.

[14] K. R. Bukkarapu and A. Krishnasamy, "Investigations on the applicability of machine learning algorithms to optimize biodiesel composition for improved engine fuel properties," *International Journal of Engine Research*, vol. 25, no. 7, pp. 1299–1314, 2024, doi: https://doi.org/10.1177/14680874241227540.

[15] S. O. Giwa and R. T. Taziwa, "Predicting fuel properties of nano-doped biodiesel using machine learning algorithms," *International Journal of Ambient Energy*, vol. 46, no. 1, 2025, doi: https://doi.org/10.1080/01430750.2025.2505064.

[16] M. Suvarna *et al.*, "Predicting biodiesel properties and its optimal fatty acid profile via explainable machine learning," *Renewable Energy*, vol. 189, pp. 245–258, 2022, doi: https://doi.org/10.1016/j.renene.2022.02.124.

[17] A. N. Ibrahim, D. Q. Nada, R. Nurdiansyah, and A. Andoko, "Optimization of XGBoost hyperparameters using three-dimensional learning AVOA for retail demand prediction," *J. Tek. Ind. J. Keilmuan dan Apl. Tek. Ind.*, vol. 28, no. 1, pp. 1–12, Mar. 2026, doi: https://doi.org/10.9744/jti.28.1.1-12.

[18] S. I. Mohammad *et al.*, "Thermophysical properties of used frying oil biodiesels blended with alcohols: Robust machine learning frameworks for density prediction," *Industrial Crops and Products*, vol. 236, 2025, doi: https://doi.org/10.1016/j.indcrop.2025.121916.

[19] A. P. Rifai, S. Briliananda, and H. Aoyama, "Tool condition monitoring with convolutional neural network for milling tools and turning inserts," *J. Tek. Ind. J. Keilmuan dan Apl. Tek. Ind.*, vol. 25, no. 1, pp. 1–16, Apr. 2023, doi: https://doi.org/10.9744/jti.25.1.1-16.

[20] H. Hasse, S. Schmitt, and F. Jirasek, "Artificial intelligence in thermodynamics: Hybrid modeling of thermophysical properties of fluids," *Current Opinion in Chemical Engineering*, vol. 51, 2026, doi: https://doi.org/10.1016/j.coche.2026.101236.

[21] F. Jirasek and H. Hasse, "Combining machine learning with physical knowledge in thermodynamic modeling of fluid mixtures," *Annual Review of Chemical and Biomolecular Engineering*, vol. 14, pp. 31–51, 2023, doi: https://doi.org/10.1146/annurev-chembioeng-092220-025342.

[22] K. R. Bukkarapu and A. Krishnasamy, "A critical review on available models to predict engine fuel properties of biodiesel," *Renewable and Sustainable Energy Reviews*, vol. 155, 2022, doi: https://doi.org/10.1016/j.rser.2021.111925.

[23] M. Alruqi, P. Sharma, B. Deepanraj, and F. Shaik, "Renewable energy approach towards powering the CI engine with ternary blends of algal biodiesel-diesel-diethyl ether: Bayesian optimized Gaussian process regression for modeling-optimization," *Fuel*, vol. 334, 2023, doi: https://doi.org/10.1016/j.fuel.2022.126827.

[24] B. Wang and I. Alruyemi, "Comprehensive modeling in predicting biodiesel density using Gaussian process regression approach," *BioMed Research International*, vol. 2021, 2021, doi: https://doi.org/10.1155/2021/6069010.

[25] G. Yang, Q. Xu, N. Liu, W. Wu, M. Zhou, and X. Wang, "Accurate prediction of thermodynamic properties of R1234yf refrigerant using Gaussian process regression models," *International Journal of Refrigeration*, vol. 186, 2026, doi: https://doi.org/10.1016/j.ijrefrig.2026.106907.

[26] M. Zhou, C. Ni, and X. Wang, "Data-driven modeling of thermophysical properties using Gaussian process regression," presented at the AIAA SciTech Forum and Exposition, 2024, 2024. doi: https://doi.org/10.2514/6.2024-0869.

[27] J. J. Cano-Gómez, G. A. Iglesias-Silva, P. Rivas, C. O. Díaz-Ovalle, and F. De Jesús Cerino-Córdova, "Densities and Viscosities for Binary Liquid Mixtures of Biodiesel + 1-Butanol, + Isobutyl Alcohol, or + 2-Butanol from 293.15 to 333.15 K at 0.1 MPa," *Journal of Chemical & Engineering Data*, vol. 62, no. 10, pp. 3391–3400, Oct. 2017, doi: https://doi.org/10.1021/acs.jced.7b00440.

[28] R. Ait Belale, F. E. M'hamdi Alaoui, Y. Chhiti, A. Sahibeddine, N. Munoz Rujas, and F. Aguilar, "Study on the thermophysical properties of waste cooking oil biodiesel fuel blends with 1-butanol," *Fuel*, vol. 287, p. 119540, Mar. 2021, doi: https://doi.org/10.1016/j.fuel.2020.119540.

[29] R. Aitbelale, M. Achak, Y. Chhiti, N. Mŭnoz Rujas, F. Aguilar, and F. E. M'hamdi Alaoui, "Effects of 2-Butanol addition on waste cooking oil biodiesel density: An updated experimental measurement and thermodynamic modeling study," *Journal of Chemical & Engineering Data*, vol. 66, no. 8, pp. 3123–3141, Aug. 2021, doi: https://doi.org/10.1021/acs.jced.1c00220.

[30] A. Sumayli, "Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil: Gaussian process regression (GPR), multilayer perceptron (MLP), and K-nearest neighbor (KNN) regression models," *Arabian Journal of Chemistry*, vol. 16, no. 7, p. 104833, Jul. 2023, doi: https://doi.org/10.1016/j.arabjc.2023.104833.

[31] A. Sumayli and S. M. Alshahrani, "Modeling and prediction of biodiesel production by using different artificial intelligence methods: Multi-layer perceptron (MLP), Gradient boosting (GB), and Gaussian process regression (GPR)," *Arabian Journal of Chemistry*, vol. 16, no. 7, p. 104801, Jul. 2023, doi: https://doi.org/10.1016/j.arabjc.2023.104801.

[32] N. Sultana, S. M. Z. Hossain, M. Abusaad, N. Alanbar, Y. Senan, and S. A. Razzak, "Prediction of biodiesel production from microalgal oil using Bayesian optimization algorithm-based machine learning approaches," *Fuel*, vol. 309, p. 122184, Feb. 2022, doi: https://doi.org/10.1016/j.fuel.2021.122184.

[33] I. Yahya, M. Gassoumi, M. Attia, and R. Ennetta, "Application of AI and machine learning techniques to improve microwave-assisted biodiesel yield prediction," *Biomass and Bioenergy*, vol. 202, p. 108250, Nov. 2025, doi: https://doi.org/10.1016/j.biombioe.2025.108250.

[34] K. Raj Bukkarapu and A. Krishnasamy, "Support vector regression approach to optimize the biodiesel composition for improved engine performance and lower exhaust emissions," *Fuel*, vol. 348, p. 128604, Sep. 2023, doi: https://doi.org/10.1016/j.fuel.2023.128604.