T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title	Bayesian optimized automated ensemble machine learning for predicting biodiesel yield from waste cooking oil: A SHAP interpretability approach
Authors	Md. Rubel, AVIAN CRIES, M.M. Harussani, Eric Kolor, Sasipa Boonyubol, Koichi Mikami, Muhammad Aziz, Jeffrey S. Cross
Citation	Chemical Engineering Research and Design, Vol. 230, , Page 985-999
Pub. date	2026, 5
DOI	https://dx.doi.org/10.1016/j.cherd.2026.05.027
Creative Commons	Information is in the article.

Powered by T2R2 (Science Tokyo Research Repository)

Chemical Engineering Research and Design 230 (2026) 985–999

Contents lists available at ScienceDirect

Chemical Engineering Research and Design

journal homepage: www.elsevier.com/locate/cherd

Bayesian optimized automated ensemble machine learning for predicting biodiesel yield from waste cooking oil: A SHAP interpretability approach

Md. Rubel [a,*], Cries Avian [b], M.M. Harussani [a], Eric Kolor [a], Sasipa Boonyubol [a], Koichi Mikami [a], Muhammad Aziz [c], Jeffrey S. Cross [a,*]

[a] Department of Transdisciplinary Science and Engineering, School of Environment and Society, Institute of Science Tokyo, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
[b] Department of Electrical Engineering, Universitas Brawijaya Malang, Jawa Timur 65145, Indonesia
[c] Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan

ARTICLE INFO

Keywords:
Waste cooking oil (WCO)
Automated ensemble learning
Process interaction mapping
SHAP interpretability
Two-step biodiesel production
Thermo-chemical boundaries

ABSTRACT

Biodiesel production from waste cooking oil (WCO) is fundamentally constrained by complex, nonlinear interactions among transesterification parameters, which conventional linear optimization fails to capture. This study addresses this knowledge gap by developing an automated ensemble machine learning (AutoML) framework within the PyCaret environment. Utilizing 49 in-house experimental data points derived from a previously established two-step acid—base catalysis process, 25 machine learning (ML) regression algorithms were systematically benchmarked. The top 15 models were optimized via Optuna's Tree-structured Parzen Estimator (TPE) Bayesian method to construct bagging, boosting, and stacking ensemble strategies. The boosting ensemble model served as a process interaction mapping framework, achieving superior performance under 3-fold cross-validation (R^2 = 0.987, MAE = 0.599, and RMSE = 0.836), significantly outperforming individual learners as well as bagging and stacking ensemble architectures. Shapley Additive exPlanations (SHAP) analysis served as a sophisticated tool for operational boundary identification, deconstructing model sensitivity regarding reaction temperature and time. Furthermore, external validation using 123 literature-derived points demonstrated significant predictive capability (R^2 = 0.973), proving the boosting ensemble's promising generalizability across related WCO-based transesterification datasets. This framework establishes an operational space mapping protocol, identifying biodiesel yield and optimal transesterification conditions through data-driven modeling.

1. Introduction

The growing global dependence on fossil fuels and the associated environmental burdens have accelerated the search for renewable and sustainable energy alternatives (Sultana et al., 2022; Lyman, 2025; Bilanovic et al., 2009; Agrawal et al., 2024). To achieve carbon neutrality, a transition from nonrenewable energy sources to sustainable energy resources, such as nuclear, wind, biomass conversion, tidal, and solar energy, is imperative (Agrawal and Rao, 2012, 2014, 2021). Among these methods, biomass conversion is particularly effective for transforming waste into valuable biofuels (Dhawane and Halder, 2019), including biodiesel (Mathew et al., 2021), bioethanol (Chen et al., 2021), biochar (Wang and Wang, 2019), and biogas (Mulu et al., 2021), which are regarded as sustainable solutions to global energy challenges (Gupte et al., 2022; Huang et al., 2022a; Katongtung et al., 2022). Biodiesel, in particular, stands out for its biodegradability, low toxicity, renewability, and reduced sulfur and CO_2 emissions (Khan et al., 2020; Najaf-Abadi et al., 2024; Paryanto et al., 2019; Thushari and Babel, 2018; Baskar et al., 2018; Gonçalves et al., 2024; Hong et al., 2016). Moreover, most modern diesel engines can operate efficiently on blends containing up to 20% biodiesel without major modifications, facilitating its integration into existing fuel infrastructures (Bhatia et al., 2021).

Biodiesel consists mainly of fatty acid alkyl esters and is typically synthesized by the transesterification of triglycerides or the esterification of long-chain fatty acids (Agrawal et al., 2024). A two-step acid—base catalysis process offers advantages over conventional single-step base catalysis for high free fatty acid (FFA) feedstock such as WCO, including higher biodiesel yield, reduced soap formation, and mitigation of corrosion caused by residual FFAs (Ali and Fadhil, 2013; Maddikeri

* Corresponding authors.
E-mail addresses: rubel.m.aa@m.titech.ac.jp (Md. Rubel), office-cross@tse.ens.titech.ac.jp (J.S. Cross).

https://doi.org/10.1016/j.cherd.2026.05.027
Received 2 February 2026; Received in revised form 8 May 2026; Accepted 16 May 2026
Available online 21 May 2026
0263-8762/© 2026 The Author(s). Published by Elsevier Ltd on behalf of Institution of Chemical Engineers. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

et al., 2012). However, feedstock costs remain the dominant economic barrier, accounting for approximately 60%–80% of the total biodiesel production costs (Dhawane et al., 2016a). Although edible oils (Altikriti et al., 2015) are widely used as feedstocks, non-edible oils (Fadhil et al., 2020; Hassan and Fadhil, 2025), waste oils (Girish et al., 2013; Gouran et al., 2021; Mohadesi et al., 2022), and microalgae (Sultana et al., 2022) provide more sustainable alternatives. In Japan, the low self-sufficiency of edible oils makes their use for biodiesel production particularly unsustainable (Cai et al., 2015; Statista Inc., 2023; Parija, 2022), whereas approximately 400,000 tons of waste cooking oil (WCO) are discarded annually, representing an abundant, low-cost, and sustainable feedstock that supports a circular economy and reduces overall production costs (Cai et al., 2015; The Asahi Shimbun Company, 2024; India's WCO, 2023). Additionally, recent advancements have further emphasized the role of WCO valorization in achieving high-efficiency biodiesel production while minimizing environmental impacts through a circular economy approach (Wang et al., 2025). Consequently, this study employs WCO as the feedstock and a two-step acid-base catalysis process for biodiesel production.

Moreover, recent technological advancements have focused on reducing biodiesel production costs through process optimization (Dhawane et al., 2016b). In alkaline transesterification, the biodiesel yield is governed by complex, nonlinear interactions among key operational parameters, such as reaction time, temperature, oil-to-methanol (MeOH) molar ratio, and catalyst loading (Bastos et al., 2020; Mairizal et al., 2020; Moradi et al., 2013). Traditional optimization approaches, including response surface methodology (RSM) and full-factorial experimental design, often struggle to adequately capture these nonlinear relationships and interaction effects (Feng et al., 2021). Prior studies have therefore relied either on generalized predictive model analyses across multiple feedstocks or on an n^m factorial design (where n is the number of levels and m is the number of factors) for a single feedstock (Sebayang et al., 2023a, 2023b; Silitonga et al., 2020; Kodgire et al., 2023; Liu et al., 2023). However, as noted by Stamenković et al., these methods lack flexibility and predictive robustness (Stamenković et al., 2013). Consequently, there is a fundamental need for advanced computational tools that go beyond simple yield prediction to offer deep interpretative insights into reaction kinetics. Machine learning (ML) offers powerful tools for modeling nonlinear dependencies and building accurate predictive models from multidimensional datasets. By extracting patterns from experimental data, ML enables efficient process optimization with minimal experimental effort (Agrawal et al., 2024). ML has gained increasing traction across molecular and materials science (Butler et al., 2018; Toyao et al., 2019), chemical engineering (Schweidtmann et al., 2021), biological processes (Greener et al., 2022), and energy conversion (Ascher et al., 2022), with emerging applications in diverse domains (Reichstein et al., 2019; Zhou et al., 2022). In the context of biodiesel research, ML has been used to model and optimize transesterification processes, thereby reducing the experimental burden and improving the predictive capability (Elmaz et al., 2020; Ma et al., 2021; Zhang et al., 2023a; Huang et al., 2022b). In the specific context of WCO-based biodiesel production, a growing body of literature has demonstrated the scope of diverse ML architectures for biodiesel yield prediction and process optimization. Dharmalingam et al. (Dharmalingam et al., 2023) demonstrated that neural network models outperform RSM in optimizing biodiesel production from mixed WCO using heterogeneous biocatalysts, while Azhar et al. (Azhar et al., 2025) employed artificial intelligence (AI)-driven modeling with particle swarm optimization (PSO) for fats, oils and grease (FOG)-based biodiesel production, highlighting the growing role of intelligent optimization in biofuel systems (Azhar et al., 2025). Recently, Zakir Hossain et al. (Zakir Hossain et al., 2022) applied a super learner ensemble approach Bayesian Optimization Algorithm-Support Vector Regression (BOA-SVR), Bayesian Optimization Algorithm-Boosted Regression Tree (BOA-BRT) for biodiesel synthesis prediction and optimization, reporting strong generalizability across diverse feedstock conditions. Nevertheless, challenges persist due to the diversity of feedstocks, variability in operational conditions, and the complexity of simultaneously optimizing multiple process parameters (Abusweireh et al., 2022; Jayaprabakar et al., 2019; Zhang et al., 2023b; Tang et al., 2015). Additionally, recent studies have highlighted the transition toward advanced heterogeneous systems for biodiesel production, such as graphene-based catalysts, which offer superior catalytic activity but introduce relatively high synthesis costs (Nazloo et al., 2023; Dong et al., 2025).

Previous studies on biodiesel yield prediction from WCO have used primarily individual ML algorithms, with limited emphasis on systematic benchmarking and advanced ensemble strategies (Ahmad et al., 2023; Buasri et al., 2023; Almohana et al., 2022). Specifically, Ahmad et al. (Ahmad et al., 2023) combined gradient boosting, extreme gradient boosting (XGBoost), and light gradient boosting machine (LGBM) models with a genetic algorithm (GA) to optimize the biodiesel yield from WCO, while Buasri et al. (Buasri et al., 2023) utilized an artificial neural network (ANN) within microwave-assisted reactors. Furthermore, Almohana et al. (Almohana et al., 2022) applied ensemble methods such as AdaBoost with Huber regression, decision trees, and Gaussian processes; however, these studies did not exploit automated ML frameworks nor did they provide the extensive external validation presented in this work.

Existing literature frequently lacks systematic benchmarking across a broad set of algorithms, automated ensemble integration, comprehensive interpretability analysis, and rigorous external validation, leaving a gap in the understanding of the thermo-chemical boundaries governing waste-to-energy conversion. To address these limitations and maximize the energy recovery potential of WCO as a high-value bioresource, building upon the experimental foundation established in a previous study (Process-I) (Rubel et al., 2026), this study develops a reproducible automated ensemble ML framework to elucidate the complex, non-linear kinetic landscape of WCO transesterification. Unlike previous manually optimized or single-model approaches, this workflow systematically benchmarks 25 distinct regression algorithms within the AutoML PyCaret (Ali, 2020a, 2020b) environment to identify the most robust and generalizable model rather than relying on an arbitrary choice; subsequently, the top 15 models were optimized via Optuna's TPE Bayesian method and incorporated into bagging, boosting, and stacking ensemble architectures. Model robustness was assessed through multiple k-fold ($k = 3$, 5, and 10) cross-validation (CV) schemes, achieving significant predictive fidelity, which was further validated against an extensive independent set of 123 data points derived from published literature records. Crucially, the decision-making logic of the finalized ensemble was deconstructed via Shapley Additive exPlanations (SHAP) analysis to establish a formal kinetic interaction mapping. This framework provides a computational deconstruction of the trained model's response surface, quantifying how process variables synergistically influence WCO conversion into fatty acid methyl esters (FAME). By identifying the critical thermo-chemical boundaries and the onset of thermal inhibition, this work defines interpretable operational boundaries for WCO-based biodiesel yield prediction within the investigated two-step acid-base catalysis process, without proposing new reaction mechanisms. Therefore, this study advances prior research by integrating automated model selection, multiple ensemble strategies, and rigorous external validation within a unified framework.

2. Materials and methods

The overall workflow of the automated ensemble ML framework, from data generation and ensemble construction to model interpretation and external validation, is illustrated in Fig. 1.

2.1. Biodiesel synthesis and experimental data collection

The WCO feedstock utilized in this study was sourced from a local

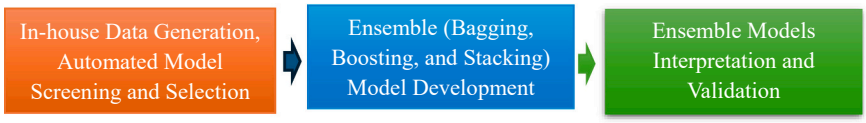

Fig. 1. Workflow of the automated ensemble machine learning system for biodiesel yield prediction.

restaurant in Tokyo, Japan, and was characterized by an initial FFA content of 10.7%. To mitigate the negative impacts of moisture (0.09%) and solid impurities, which are known to cause catalyst deactivation (Zhu et al., 2024) and promote hydrolysis, the WCO underwent a sequential four-step filtration process followed by thermal drying at 100–110 °C for 1 h (Rubel et al., 2026). Furthermore, an acid-catalyzed esterification step was conducted to reduce the FFA content and ensure successful deacidification, thereby preventing saponification during the subsequent alkaline stage. This protocol effectively lowered the FFA level to 0.89% and the acid value (AV) to 1.78 mg KOH / g, validating the suitability of the pre-esterified sample for the base-catalyzed transesterification phase established in a previous study (Process-I) (Rubel et al., 2026).

Briefly, the optimized transesterification involved converting triglycerides into FAME using a KOH catalyst. To ensure high separation efficiency, the post-reaction mixture was transferred to a separatory funnel and allowed to settle for 1 h to facilitate complete phase separation between the FAME and glycerol layers. The bottom glycerol layer was discarded, and the isolated biodiesel phase was washed and dried to remove residual contaminants and moisture (Rubel et al., 2026). To explore the fundamental thermochemical boundaries of the system, the transesterification phase was systematically investigated by varying four critical process parameters: reaction time (15–90 min), reaction temperature (35–60 °C), oil-to-methanol molar ratio (1:3–1:8), and catalyst loading (1–4 wt%). Additionally, to ensure consistent mass transfer and eliminate mixing-induced variability, the stirring speed was maintained constant at 600 revolutions per minute (RPM) for all 49 experimental trials. Consequently, RPM was excluded as an input feature in the ML models due to its zero variance in this specific dataset. However, it is recognized that varying the stirring speed can significantly influence reaction rates and alter the non-linear interaction between parameters (Prajapati et al., 2024; Yuan et al., 2024a). The resulting biodiesel compositions were quantified via Gas Chromatography-Mass Spectrometry (GC—MS) using a calibrated internal standard method. Following established protocols, the peak area percentages were correlated with FAME concentrations via multi-point calibration curves (0.5–4 mg/ml) (Usman et al., 2023). The biodiesel yield percentage was then calculated according to Eq. (1), which incorporates both the gravimetric mass of the synthesized product and the calibrated FAME purity to ensure precise quantification for the ML dataset (Ahmad et al., 2023).

$$\text{Biodiesel yield}(\%) = \frac{\text{Total biodiesel produced}(\text{gm}) \quad * \quad (\%\text{FAME from GC} - \text{MS})}{\text{Total oil used in reaction}(\text{gm})} \text{x}100 \tag{1}$$

This experimental design generated 49 discrete data points from the primary WCO valorization process (Process-I), providing the high-quality dataset required for the development of the automated ensemble framework (Annexure I). To assess whether the model captures transferable process trends rather than only lab-specific patterns, an additional independent dataset of 123 biodiesel yield records was compiled from five peer-reviewed studies for external validation and model generalizability (Annexure II). Although the primary experimental dataset comprises 49 data points, the risk of overfitting was mitigated by using ensemble techniques that aggregate multiple tuned base learners, which reduces variance and improves robustness for small-to-moderate sample sizes, together with a k-fold CV strategy to provide reliable performance estimates. Furthermore, external validation using 123 literature-derived data points provides a critical assessment of model generalizability beyond the limited in-house data points.

2.2. Data preprocessing

The data sets utilized in this study were verified for completeness with no missing values identified across the experimental records. A direct-feature architecture was adopted to preserve the physical significance of the process variables. This approach ensures that all operational parameters — reaction temperature, reaction time, oil-to-MeOH molar ratio, and catalyst loading—contribute to the model objective function based on their original experimental scales, thereby preserving the kinetic relevance of the input data (Ahsan et al., 2021; Kim et al., 2025).

2.3. Machine learning modeling framework

The ML modeling framework is summarized in Fig. 2 and comprises dataset preparation, regression benchmarking, hyperparameter optimization, ensemble construction, and model interpretation and validation. Initially, the experimental data were organized into input features (reaction time, reaction temperature, oil-to-methanol molar ratio, and catalyst loading) and the target variable (biodiesel yield), with the RPM feature excluded to prioritize primary operational drivers. Benchmarking was performed on 25 regression algorithms using the AutoML PyCaret library, where performance was assessed across 3-, 5-, and 10-fold CV schemes.

The 15 best-performing models, selected based on R^2, MAE, and RMSE, subsequently underwent Bayesian hyperparameter optimization via Optuna's TPE algorithm before ensemble synthesis. These optimized learners served as base models for bagging, boosting, and stacking ensemble architectures. Bagging was implemented to reduce variance, boosting to decrease bias through sequential learning, and stacking to integrate complementary predictions via a meta-model. Final model interpretability was assessed using SHAP analysis to quantify the influences of the process parameters, while generalizability was rigorously validated against an independent, literature-derived dataset.

2.3.1. Automated model benchmarking and selection

Initial model benchmarking was performed via the AutoML PyCaret library, which evaluated 25 regression algorithms under a unified preprocessing pipeline utilizing multi-level k-fold (k = 3, 5, and 10) CV. The use of 25 algorithms reflects the complete regression candidate pool available within the PyCaret AutoML library (Ali, 2020a, 2020b), ensuring unbiased, data-driven model selection across diverse algorithmic families rather than arbitrary pre-selection. The candidate pool spanned a comprehensive array of computational architectures,

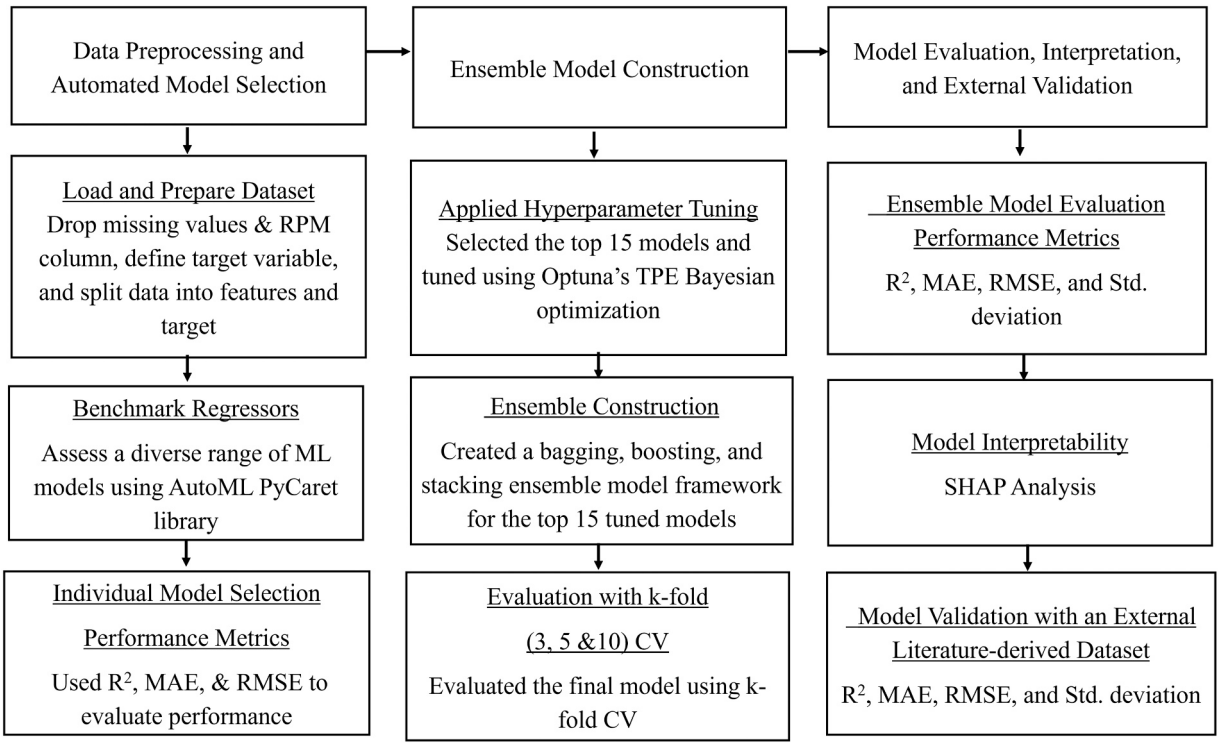

Fig. 2. Schematic of the machine learning pipeline used for biodiesel yield prediction.

including: (a) linear and regularized regressors: linear regression (LR), ridge regression (Ridge), elastic net (EN), least absolute shrinkage and selection operator (Lasso), least angle regression (LAR), Lasso least angle regression (LassoLARS, LLAR), and orthogonal matching pursuit (OMP); (b) Bayesian and robust learners: Bayesian Ridge (BR), automatic relevance determination (ARD), Huber regressor, and random sample consensus (RANSAC); (c) tree-based and ensemble architectures: decision tree (DT), random forest (RF), extra trees (ET), and gradient boosting regressor (GBR); (d) boosting and advanced ensembles: adaptive boosting (AdaBoost), categorical boosting (CatBoost), light gradient boosting machine (LightGBM), and extreme gradient boosting (XGBoost); (e) instance- and kernel-based models: K-Nearest Neighbors (KNN), support vector machine (SVM), kernel ridge (KR); (f) neural and non-parametric models: multilayer perceptron (MLP), passive aggressive regressor (PAR), and the Theil-Sen robust regressor (TR). The algorithms were rigorously ranked based on their cross-validated coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). To ensure a high-fidelity candidate pool for the final ensemble construction, the top 15 performing algorithms were retained for subsequent Bayesian hyperparameter optimization. This systematic benchmarking protocol eliminates manual selection bias and establishes a robust foundation for capturing the complex non-linearities of the transesterification process, with full numerical results reported in Annexure III.

2.3.2. *Hyperparameter optimization*

The top 15 candidate models identified during benchmarking were rigorously refined via Bayesian hyperparameter optimization utilizing Optuna's TPE algorithm. The objective function was configured to maximize the cross-validated R^2, enabling an efficient exploration of high-dimensional hyperparameter spaces by concentrating evaluations in high-performance regions. Unlike static search methods, the TPE algorithm models the probabilistic distribution of hyperparameters relative to model performance, thereby prioritizing configurations with the highest likelihood of enhancing predictive accuracy (Watanabe, 2023; Sieradzki and Mańdziuk, 2025; Rong et al., 2021; Snoek et al., 2012).

In contrast, conventional grid search evaluates all discrete combinations within a predefined hyperparameter grid and can identify the optimum only if it lies inside the specified search space, which makes it computationally expensive for complex models and a wide search range (Agrawal et al., 2024). As an alternative, random search is less costly and may outperform grid search in high-dimensional settings, but it lacks the probabilistic memory and performance-guided trajectory provided by Optuna's TPE algorithm (Nishio et al., 2018; Tao et al., 2022). Furthermore, although metaheuristic methods, such as the whale optimization algorithm (WOA), offer gradient-free exploration, they often lack the systematic probabilistic modeling required for robust hyperparameter refinement in diverse model families (Agrawal et al., 2024). The hyperparameter search space for each model family is summarized in Table S1 (Annexure IV(a)), and a schematic of Optuna's TPE optimization workflow integrated with k-fold CV is provided in Fig. S1 **(Annexure IV).**

2.3.3. *Ensemble model aggregation*

To transcend the predictive limitations of individual algorithms, three distinct ensemble strategies (bagging, boosting, and stacking) were constructed from the tuned top 15 base models. Ensemble learning combines multiple ML algorithms to enhance the predictive performance beyond that of individual models by aggregating their outputs through mathematical averaging, sequential error weighting, or meta-algorithmic integration (Asadi and Hajj, 2024). Bagging (Bootstrap Aggregating) was employed via a VotingRegressor (VR) architecture to reduce variance by training base models in parallel on bootstrapped data subsets, with their outputs aggregated through averaging to achieve a more robust consensus (Breiman, 1996; Freund and Schapire, 1997; Wolpert, 1992). Conversely, boosting was employed to minimize predictive bias through a sequential learning protocol. Specifically, an AdaBoost framework was utilized with an optimized KNN base model, focusing each successive iteration on correcting the residuals of its predecessors (Freund and Schapire, 1997). Finally, Stacking integrated complementary base models by using their predictions as inputs for a secondary Ridge-regression-based meta-learner, which generates the

final output and aims to improve the generalization performance (Wolpert, 1992). The resulting ensemble models were evaluated via 3-, 5-, and 10-fold CV to mitigate overfitting and ensure the robustness of the proposed high-fidelity ensemble framework (Agrawal et al., 2024; Azhar et al., 2025; Moklis et al., 2025).

2.3.4. Model evaluation, interpretation, and validation

The predictive accuracy of the ensemble frameworks was quantified using the R^2 (Eq. 2), MAE (Eq. 3), and RMSE (Eq. 4), where y_i and $\hat{y}_i$ are the measured and predicted biodiesel yields, respectively, and $\bar{y}$ is the mean of the measured yields (y_i), and n is the total number of experimental observations.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

Beyond statistical metrics, model interpretability was assessed using SHAP, a game-theoretic, model-agnostic framework implemented via a KernelExplainer architecture with the training dataset serving as the background reference to establish the baseline expected prediction (Lundberg and Lee, 2017; Lundberg et al., 2020). The SHAP framework decomposes individual model predictions into additive feature-level contributions, enabling both local and global interpretability of the finalized ensemble framework, as detailed in Section 3.5 (Agrawal et al., 2024; Lundberg and Lee, 2017).

3. Results and discussion

3.1. Feature correlation analysis

Pearson correlation coefficients were utilized to quantify the linear interdependence between the input variables (reaction times, temperature, oil-to-methanol molar ratio, and catalyst concentration) and the resulting biodiesel yield, as well as to screen for potential multicollinearity. As illustrated in the correlation heatmap (Fig. 3), the analysis revealed weak-to-moderate correlations between individual inputs and yield. Reaction times (r = 0.320) and the oil-to-MeOH ratio (r = 0.295) exhibited positive correlations with yield, while catalyst concentration (r = -0.549) and reaction temperature (r = -0.220) demonstrated negative linear trends. This trend indicates that while a baseline amount of catalyst is required to initiate the transesterification reaction, excessive alkaline loading can trigger irreversible saponification, thereby reducing the biodiesel yield (Degfie et al., 2019). Critically, the pairwise correlations among the input variables remained consistently low ($|r| < 0.25$), confirming the absence of significant multicollinearity and justifying the retention of all four features within the predictive framework. The negative correlations observed for catalyst concentration and temperature are particularly noteworthy, as they reflect the net linear trend in regions where yield decreases beyond the optimum conditions. This negative correlation for temperature, especially beyond 60 °C, is fundamentally linked to the physical properties of methanol (MeOH) and shifting reaction kinetics. As the reaction temperature approaches the boiling point of MeOH (64.7 °C), a liquid-to-vapor phase transition occurs. This evaporation reduces the effective concentration of MeOH at the oil-catalyst interface, thereby hindering the forward transesterification kinetics (Rubel et al., 2026; Yuan et al., 2024b). Furthermore, as reported in recent literature, high thermal energy can accelerate undesirable competitive side reactions, such as the saponification of triglycerides. This process consumes the alkaline catalyst and leads to soap formation, ultimately resulting in the observed decline in biodiesel yield (Nadim et al., 2025). However, the relatively low magnitude of these coefficients underscores the inherent limitations of Pearson's *r* in capturing the complex, non-linear dependencies known to govern the transesterification kinetics (Schober et al., 2018; Lin et al., 2022). This mathematical gap necessitates the application of the advanced ensemble architecture developed in this study, which is specifically designed to deconstruct these high-dimensional, non-linear chemical interactions.

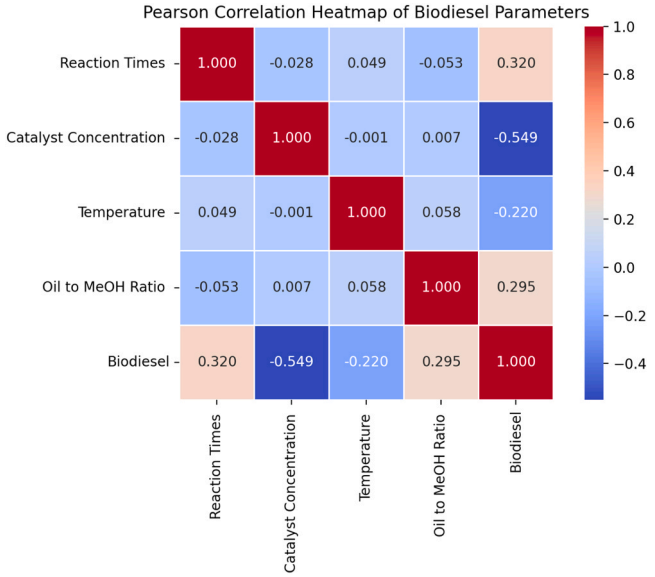

Fig. 3. Pearson correlation matrix for the input variables and biodiesel yield.

3.2. Individual model benchmarking and Bayesian optimization

Initial benchmarking of the 25 ML models under default configurations identified the ET regressor as the premier baseline performer. Furthermore, several models displayed negative R^2 values, as illustrated in Fig. 4. These negative R^2 scores indicate that the model's predictive accuracy was lower than a simple horizontal line representing the mean of the observed yields, meaning the models failed to capture the underlying data patterns (Chicco et al., 2021). However, the subsequent application of Bayesian hyperparameter optimization via Optuna's TPE algorithm induced a significant shift in the relative model rankings (Fig. 4). The most notable performance trajectory was observed in the KNN regressor, which exhibited a substantial elevation in predictive accuracy, with its cross-validated R^2 value increasing from approximately 0.305 (pre-tuning) to a top-performing score of 0.689 (post-tuning under 5-fold CV). The sensitivity of the KNN architecture to localized data density necessitated rigorous evaluation across 3-, 5-, and 10-fold CV schemes to ensure numerical stability. Following the optimization phase, the 15 best-performing models were retained as a diverse candidate pool to provide complementary predictive behavior for the subsequent ensemble construction (Fig. 4). The finalized R^2 values for the selected architectures were: KNN ($R^2 = 0.689$), ET ($R^2 = 0.617$), AdaBoost ($R^2 = 0.321$), OMP ($R^2 = 0.464$), LR ($R^2 = 0.464$), LAR ($R^2 = 0.464$), LLAR ($R^2 = 0.464$), GBR ($R^2 = 0.216$), ARD ($R^2 = 0.437$), RANSAC ($R^2 = 0.396$), RF ($R^2 = 0.138$), Huber ($R^2 = 0.307$), DT ($R^2 = 0.181$), EN ($R^2 = 0.206$), and Ridge ($R^2 = 0.169$). These results highlight that while tree-based learners offer strong baseline performance, instance-based and regularized linear models, when properly tuned, provide the high-fidelity local error correction required for the complex transesterification landscape. The comprehensive numerical results for all 25 benchmarked models are documented in Annexure III.

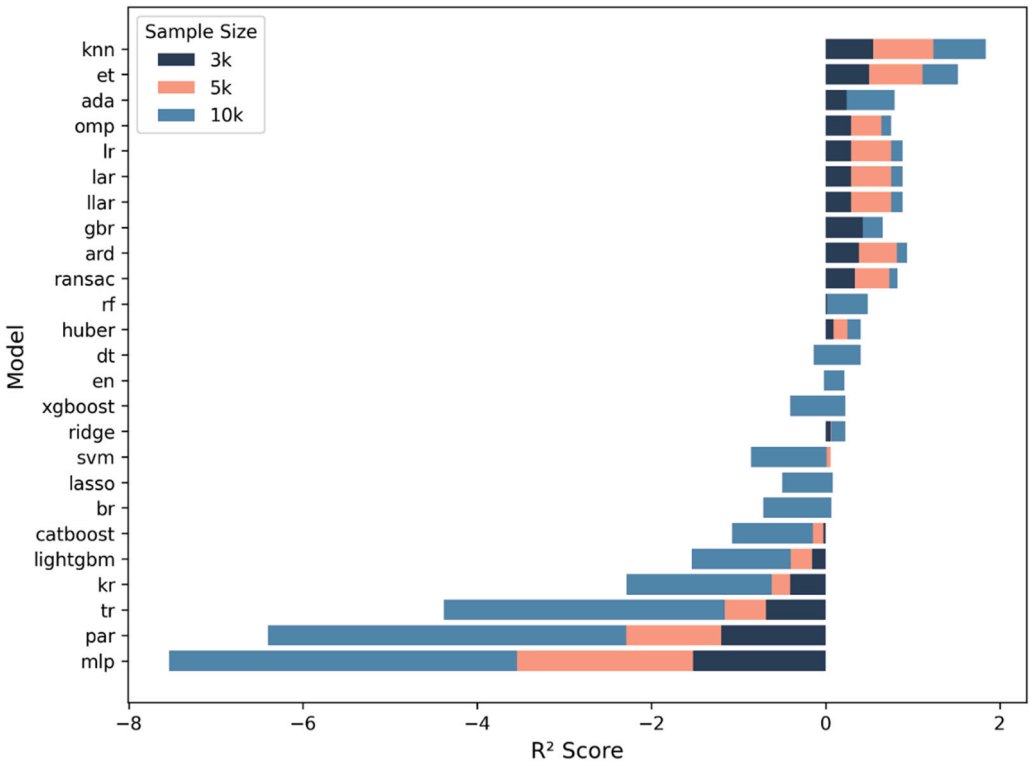

Fig. 4. Performance comparison of 25 machine learning models after hyperparameter tuning with Optuna's TPE algorithm.

3.3. Ensemble architecture optimization and performance evaluation

The predictive performance of the transesterification framework was further augmented through the construction of three ensemble architectures: bagging, boosting, and stacking. Each ensemble was rigorously optimized via Optuna's TPE Bayesian strategy and evaluated across 3-, 5-, and 10-fold CV schemes to ensure numerical stability and generalization. The key ensemble configurations and selected optimal hyperparameters for the best-performing cross-validation fold of each architecture are summarized in Table 1, with complete hyperparameter search spaces and fold-wise configurations documented in Annexure IV (b).

3.3.1. Bagging ensemble performance

From the candidate pool of 15 optimized base models (Section 3.2), bagging-type ensembles were constructed via an automated aggregation and blending pipeline. The finalized configurations were structured as VotingRegressor models, strategically combining diverse architectures to maximize predictive consensus: ET and AdaBoost for 3-fold CV; ET, AdaBoost, and KNN for 5-fold CV; and ET and KNN for 10-fold CV scheme. These specific combinations provided the optimal trade-off between local error correction and global stability across the experimental space. The 10-fold CV ensemble yielded the premier predictive fidelity, achieving an R^2 of 0.979, MAE = 0.648, and RMSE = 0.808. As illustrated in Fig. 5, predictive performance exhibited a positive trajectory as the CV folds increased from 3 to 10, with R^2 improving from 0.897 to 0.979 (left panel) and both MAE and RMSE decreasing substantially from 1.905 to 0.648 and 2.305–0.808, respectively (right panel). This indicates that a larger effective training fraction and more exhaustive validation cycles significantly enhanced the model stability, leading to a more robust fit of the transesterification kinetics (Althnian et al., 2021; Bailly et al., 2022). The mean R^2 of 0.934 across all CV schemes confirms that the bagging ensemble captured most of the variance in biodiesel yield. Furthermore, the decreasing standard deviation of the error metrics with increasing folds reflects more consistent predictions as data utilization improved. Detailed fold-wise performance metrics are documented in Fig. S2 and Fig. S3 of Annexure V.

3.3.2. Boosting ensemble performance

The candidate pool of 15 optimized base models was integrated into a boosting pipeline to facilitate a rigorous comparative analysis of ensemble strategies. The optimal boosting configuration, identified via the automated boosting and blending workflow, was an AdaBoost regressor utilizing KNN as the base estimator, which was selected as the best performer across all CV schemes (3-, 5-, and 10-fold), as detailed in Table S1.

As illustrated in Fig. 6(a), the AdaBoost-KNN ensemble attained its peak predictive fidelity at 3-fold CV, achieving an $R^2 = 0.987$ (left panel), with corresponding MAE = 0.599, and RMSE = 0.836 (right panel). This indicates an excellent degree of agreement between the predicted and measured biodiesel yields, largely due to the base learner's ability to map localized non-linearities in the reactive space. However, as the number of folds increased to 5 and 10, the R^2 declined marginally to 0.982 and 0.929 (left panel), while MAE and RMSE increased from 0.599 to 1.072 and 0.836–1.370, respectively (right panel). This transition reflects the increased complexity of maintaining high-precision local error correction across more diverse and fragmented validation splits. The observed increase in the fold-wise standard deviation, particularly for the RMSE at 10-fold CV, suggests a heightened sensitivity to residual variability. The mean R^2 of 0.966 across all CV schemes further confirms the overall stability of the boosting ensemble. Nevertheless, the predicted versus true biodiesel yield plot (Fig. 6(b)) provides a robust validation of the model reliability across both training and testing subsets. The high fidelity of the AdaBoost-KNN architecture is confirmed by the extreme statistical tightness of the data points along the 45° ideal diagonal across the experimental yield range of 60–90%. This alignment demonstrates that the Bayesian-optimized hyperparameters successfully captured the system's global trends while preserving localized kinetic resolution. Furthermore, the marginal distribution plots closely mirror the

Table 1
Key configuration of the ensemble models (best -performing cross-validation fold configuration). Complete hyperparameters search spaces are provided in Annexure IV(b).

Ensemble	Architecture	CV Fold	Base Models	Key Optimal Hyperparameters
Bagging	Voting-Regressor	10-fold	ET, KNN	**VR:** weights: None, n_jobs:-1, verbose: False, **ET:** max_depth: None, min_samples_split: 2, min_samples_leaf: 1, n_estimators: 100, max_features: 1.0, bootstrap: False, random_state: 123 **KNN:** algorithm: auto, n_neighbors:1 leaf_size: 30, metric: euclidean, p:2, weights: distance
Boosting	AdaBoost Regressor	3-fold	KNN (base estimator)	**AdaBoost:** n_estimators: 10, learning_rate: 1.0, loss: linear, random_state: 123 **KNN:** algorithm: auto, leaf_size: 30, metric: euclidean, n_neighbors: 1, p: 2, weights: uniform
Stacking	Stacking Regressor	10-fold	ET, CatBoost, Ridge, Huber, Lasso	**Stacking Regressor:** cv: 5, n_jobs: −1, passthrough: False, verbose: 0 **ET:** n_estimators: 100, max_depth: None, min_samples_split: 2, bootstrap: False **CatBoost:** depth: 6, n_estimators: 213, eta: 0.268, l2_leaf_reg: 4, **Ridge:** alpha: 0.560, fit_intercept: True, solver: auto, **Huber:** alpha: 0.322, epsilon: 1.585, max_iter: 100, **Lasso:** alpha: 0.727, max_iter: 1000 **meta_learner (Ridge):** alpha: 0.56, fit_intercept = True, solver: auto, n_jobs = -1, verbose = False, random_state: 123

experimental modes, confirming the ensemble's ability to capture primary kinetic trends without losing resolution in high- or low-yield regions. The close overlap between the training (blue) and test (orange) sets suggests no obvious overfitting within the evaluated data partitions. This confirms that the model remains robust when processing unseen data partitions. By achieving a peak R^2 of 0.987, the results highlight the base learner's specific capacity to map complex, non-linear kinetic interactions within the reactive space (Bailly et al., 2022; Raykov and DiStefano, 2025; Avian et al., 2024; Ribeiro and dos Santos Coelho, 2020). This predictive alignment establishes the empirical foundation necessary for the subsequent SHAP-based kinetic deconstruction, transitioning the research from black-box estimation toward transparent process mapping. Comprehensive configuration and validation data for the boosting ensemble are documented in Fig. S4 and Fig. S5 of Annexure VI.

3.3.3. *Stacking ensemble performance*

To maintain methodological consistency, the same pool of 15 optimized base models was evaluated within a Stacking Regressor framework. The stacking configuration combines multiple base models under a meta-learner to exploit complementary model behavior and reduce both bias and variance. The optimal stacking pipeline, obtained from the automated stacking and blending workflow, employs ET, CatBoost, Ridge, Huber regression, and Lasso regression as base models, with the stacking regressor using a Ridge regression meta-learner as the final estimator.

The stacking ensemble exhibited its best performance under 10-fold CV, with $R^2 = 0.849$, MAE = 1.388, and RMSE = 1.717 (Fig. 7, left and right panels, respectively). As illustrated in Fig. 7, model accuracy improved as the number of CV folds increased from 3 to 10, with R^2 improving from 0.502 to 0.849 (left panel) and both MAE and RMSE decreasing substantially from 4.456 to 1.388 and 5.150–1.717, respectively (right panel). The simultaneous increase in R^2 and the reduction in both MAE and RMSE indicate enhanced learning stability and feature representation with larger effective training sets (Althnian et al., 2021; Bailly et al., 2022). The mean R^2 of 0.700 across all CV schemes reflects a moderate but consistent predictive capability. The relatively high R^2 and moderate error metrics across folds suggest that the stacking ensemble achieved a reasonable bias—variance trade-off, although its performance remained lower than that of the bagging and boosting ensembles. However, this R^2 value of 0.849 at 10-fold CV implies that approximately 15.1% of the variance remains unexplained, indicating a

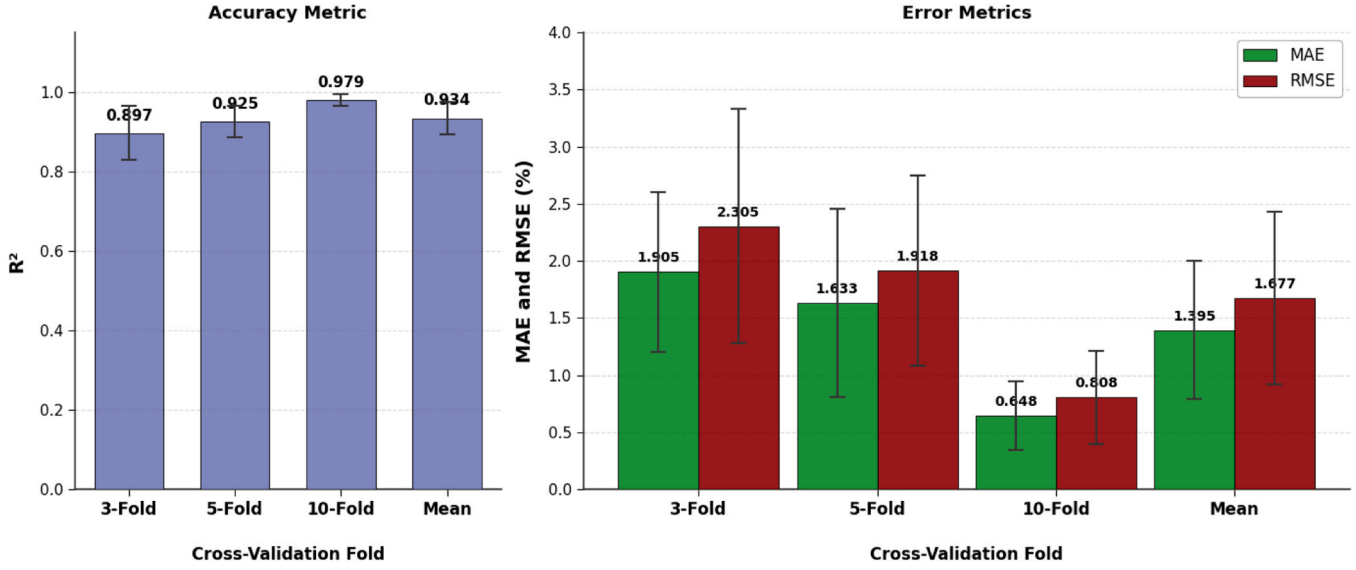

Fig. 5. Bagging ensemble performance across cross-validation folds: R^2 under 3-, 5-, and 10-fold cross-validation and their mean (left panel); MAE and RMSE (%) across the same cross-validation schemes (right panel). Error bars represent the standard deviation across folds.

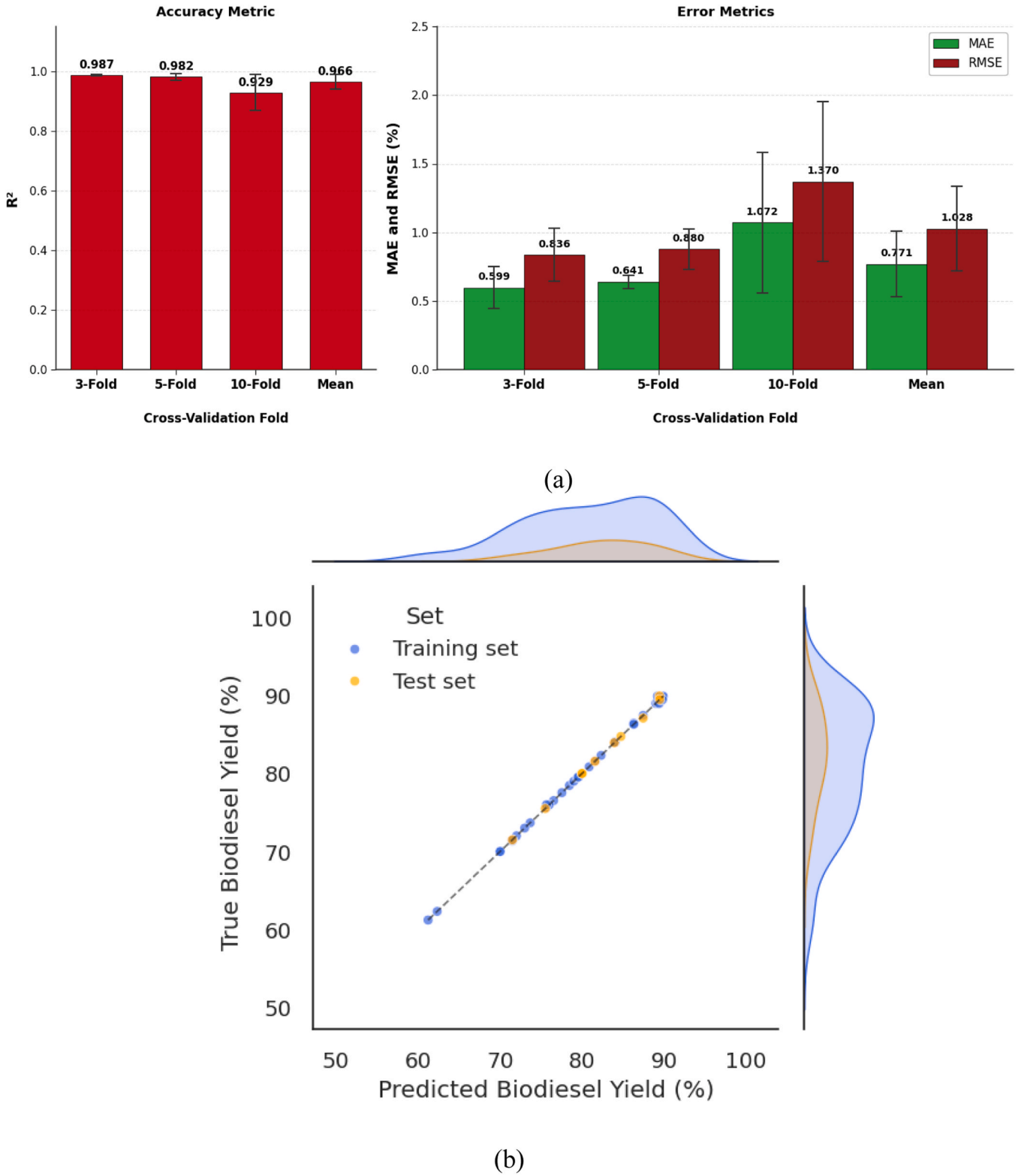

Fig. 6. Boosting ensemble: (a) R^2 (left panel) and MAE and RMSE (%) (right panel) across 3-, 5-, and 10-fold cross-validation and their mean, with error bars representing standard deviation; and (b) predicted vs. true biodiesel yield for the 3-fold cross-validation scheme.

non-negligible error margin that may stem from irreducible noise or latent variables (Wolpert, 1992). This suggests that while the stacking framework achieves a stable bias–variance trade-off, its predictive capacity is constrained compared to the more robust bagging and boosting ensembles. This highlights a limitation in its sensitivity for high-precision applications, likely due to the structural constraints of the Ridge meta-learner in effectively harmonizing the diverse predictions of the base models for this specific non-linear process. The detailed stacking configurations and fold-wise evaluation results are provided in Fig. S6 and Fig. S7 of Annexure VII.

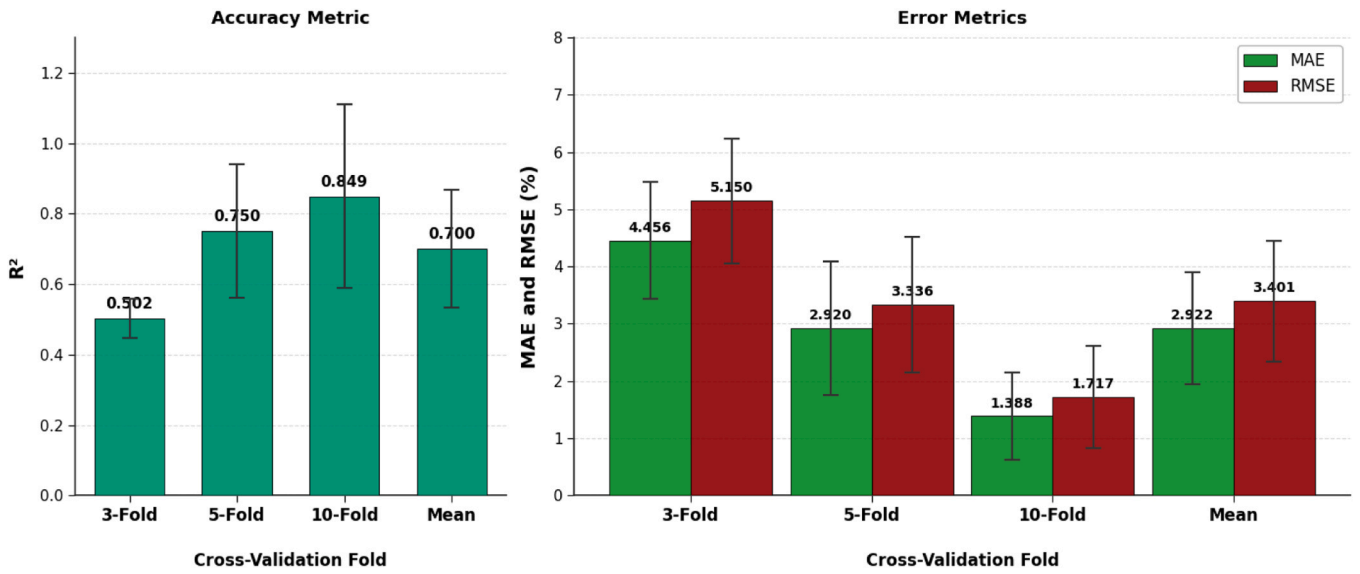

Fig. 7. Stacking ensemble: R^2 (left panel) and MAE and RMSE (%) (right panel) across 3-, 5-, and 10-fold CV and their mean, with error bars representing the standard deviation across folds.

3.4. Comparative machine learning study

Fig. 8 presents a comprehensive performance comparison of the top 15 tuned individual ML models under 5-fold CV (left panels) alongside four ML model architectures — KNN, bagging, boosting, and stacking ensembles — evaluated across 3-, 5-, and 10-fold CV schemes, as well as their calculated mean scores (right panels). The four model architectures in the right panels of Fig. 8 represent the four structural tiers of the modeling pipeline: the best individual learner after Bayesian optimization, tuned KNN, and the optimal configuration of each of the three standard ensemble architectures, namely bagging, boosting, and stacking. These ensemble approaches were selected because they operate through fundamentally distinct aggregation principles: variance reduction for bagging, bias reduction for boosting, and hierarchical meta-learning for stacking. The boosting ensemble under a 3-fold CV scheme achieved the highest predictive accuracy, with $R^2 = 0.987$, MAE = 0.599, and RMSE = 0.836. The bagging ensemble under a 10-fold CV scheme followed closely ($R^2 = 0.979$), while the stacking ensemble under a 10-fold CV scheme attained an $R^2 = 0.849$, followed by the tuned KNN model (5-fold CV), which attained an R^2 of 0.689. Detailed numerical comparisons are provided in Fig. S8 (Annexure VIII).

The superior performance of the ensemble architectures relative to the individual KNN model underscores the efficacy of aggregating multiple learners to exploit complementary predictive strengths. Specifically, boosting enhances accuracy through sequential error-focused weight updates, bagging mitigates variance by aggregating consensus from parallel resampled subsets, and stacking minimizes bias through hierarchical meta-learning. These combined mechanisms allow the ensembles to navigate the complex, non-linear kinetics of WCO transesterification with higher stability and precision than any singular algorithm. Notably, the optimized boosting ensemble demonstrates superior predictive performance by achieving the highest R^2 (0.987) and lowest MAE (0.599) among all tested architectures. Furthermore, the sensitivity of these ensemble architectures, along with the individual KNN model, to data partition density was evaluated through 3-, 5-, and 10-fold CV. For the bagging and stacking frameworks, performance improved with higher fold counts, reflecting the benefits of maximized training fractions. Conversely, the boosting ensemble achieved its optimal bias-variance equilibrium at 3-fold CV. Increasing the fold count beyond this point slightly degraded performance ($R^2 = 0.929$ at 10-fold CV), as the sequential learning process became increasingly sensitive to local residual noise within the more fragmented validation subsets, which is consistent with the tendency toward overfitting in boosted neighbor-based topologies (Althnian et al., 2021; Bailly et al., 2022; Raykov and DiStefano, 2025; Avian et al., 2024; bhagat, 2024).

As summarized in Table 2, the proposed boosting ensemble model significantly outperforms established benchmarks in the literature for WCO-to-biodiesel yield prediction. Specifically, the proposed boosting ensemble model ($R^2 = 0.987$, MAE = 0.599, RMSE = 0.836) outperforms the LGBM model reported by Ahmad et al. (Ahmad et al., 2023) ($R^2 = 0.94$, RMSE = 8.92), and the boosted Huber regressor and decision tree models ($R^2 = 0.81$ and 0.78, MAE = 3.84 and 5.94, respectively) presented by Almohana et al. (Almohana et al., 2022). Furthermore, it demonstrates superior predictive performance over the Bayesian-optimized regression trees reported by Zakir Hossain et al. (Zakir Hossain et al., 2022), which reported $R^2 = 0.81$, MAE = 8.51, and RMSE = 12.46. These comparisons highlight the effectiveness of the automated ensemble strategy and underscore the high quality and consistency of the experimental datasets utilized in this study.

3.5. Model interpretability and mechanistic insights via SHAP

To reveal the 'black-box' nature of the best boosting ensemble, SHAP analysis was employed. Rooted in cooperative game theory, this model-agnostic approach decomposes individual predictions into additive contributions from each input feature, providing both local and global interpretability. Fig. 9 presents these insights through a multi-scale lens utilizing the testing dataset, featuring a waterfall plot for local feature attribution and a beeswarm plot for global feature influence.

In the local waterfall explanation (Fig. 9(a)), the model's baseline expected yield of $E[f(X)] = 80.713$ is increased to a finalized predicted yield of $f(x) = 87.45$ through positive contributions from all four inputs: temperature (+2.00), reaction times (+2.44), oil-to-MeOH ratio (+1.25), and catalyst concentration (+1.04). This decomposition validates that at this specific operating point, reaction times and temperature are the dominant driving factors for maximizing the transesterification rate, whereas the oil-to-methanol molar ratio and catalyst concentration play supporting roles.

Conversely, the global beeswarm plot (Fig. 9(b)) provides a comprehensive mapping of feature influence, where the color gradient (red for high features values, blue for low features values) reveals the directional dynamics of each effect. Reaction temperature is identified as the primary controlling factor; while moderate temperatures are beneficial, high values (red points) are predominantly associated with

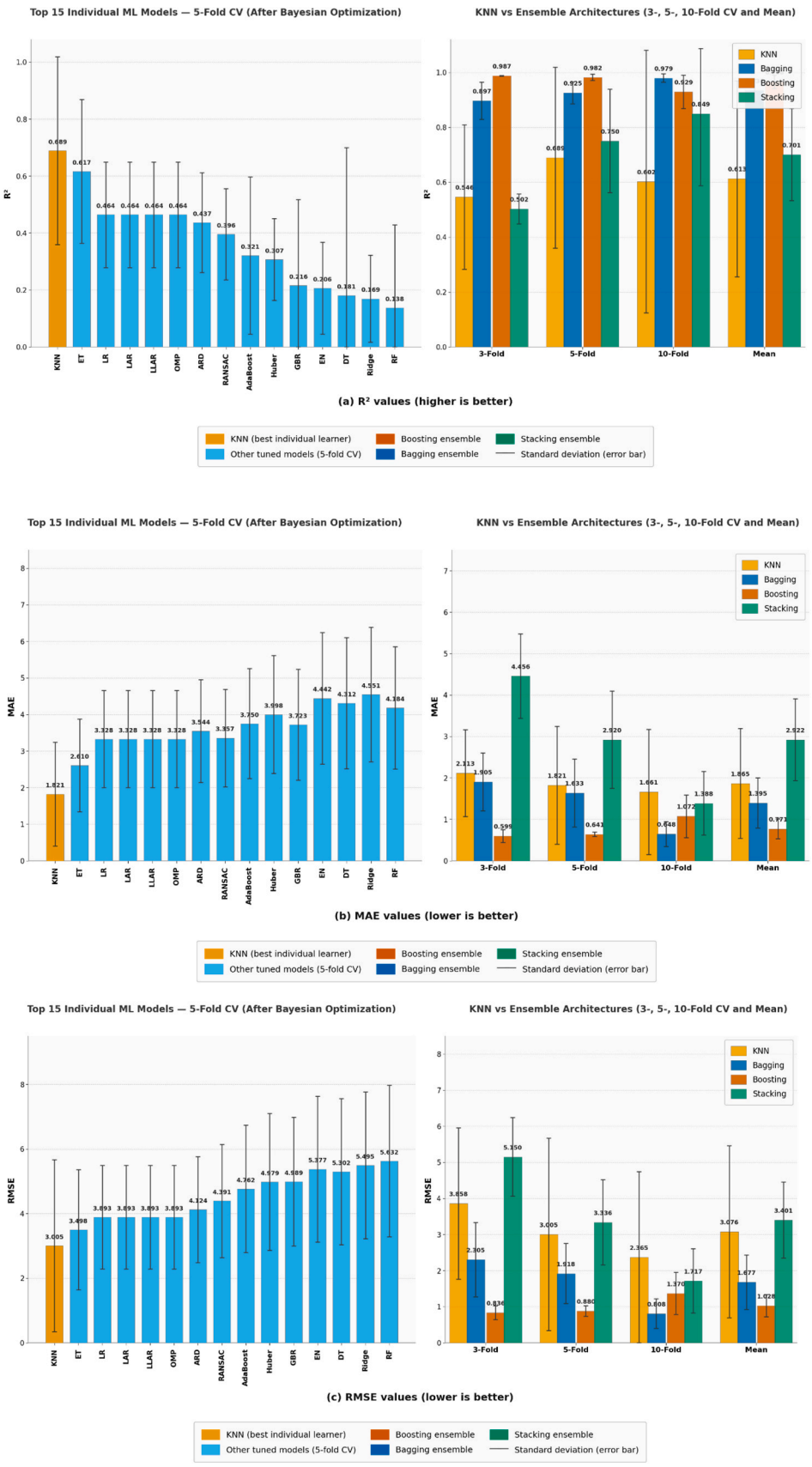

Fig. 8. Comparative performance of the top 15 tuned individual ML models under 5-fold cross-validation (left panels) and four ML model architectures — KNN, bagging, boosting, and stacking ensembles — across 3-, 5-, and 10-fold cross-validation and their mean (right panels), evaluated in terms of (a) R^2, (b) MAE, and (c) RMSE. Error bars represent the standard deviation across cross-validation folds.

Table 2
Comparative performance of the proposed ensemble model against existing literature for WCO-to-biodiesel prediction.

Author	Feedstocks	Biofuels	ML Models	R^2	MAE	RMSE
Ahmad et al (Ahmad et al., 2023).	WCO	Biodiesel	LGBM	0.94	7.72	8.92
Almohana et al (Almohana et al., 2022).	WCO	Biodiesel	AdaBoost-HBR	0.81	3.84	-
Almohana et al (Almohana et al., 2022).	WCO	Biodiesel	AdaBoost-DT	0.78	5.94	-
Zakir Hossain et al (Zakir Hossain et al., 2022).	Waste date seed oil	Biodiesel	BOA-BRT	0.81	8.51	12.46
This study	WCO	Biodiesel	KNN	0.689	1.821	3.004
			Bagging ensemble	0.979	0.648	0.808
			Boosting Ensemble	0.987	0.599	0.836
			Stacking ensemble	0.849	1.388	1.717

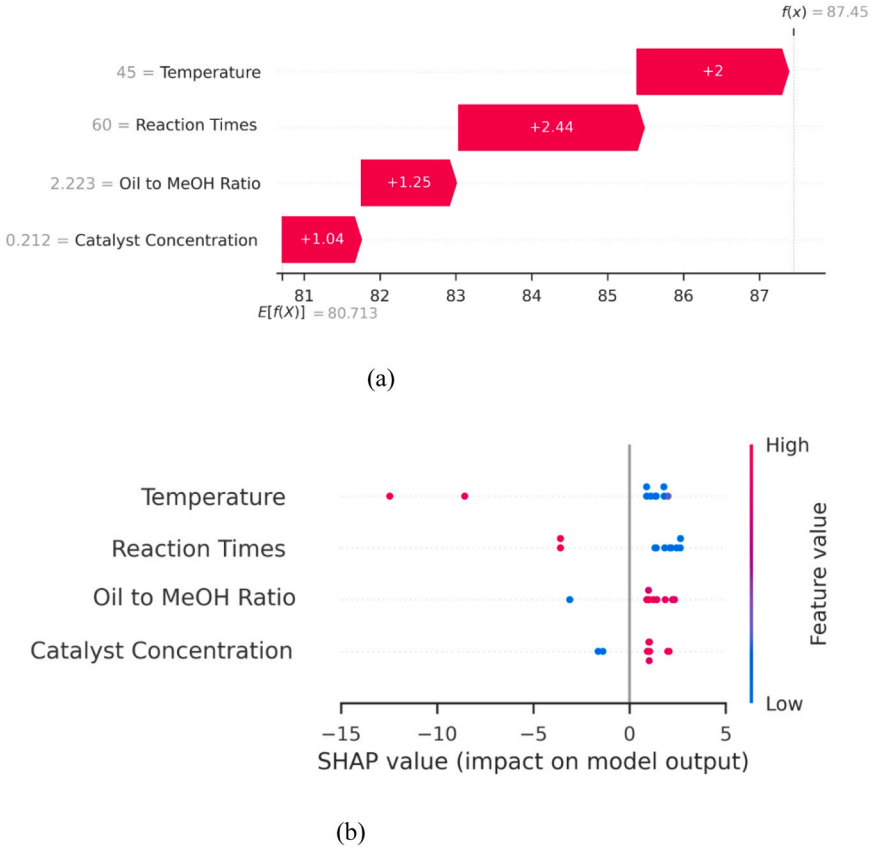

Fig. 9. SHAP analysis of the biodiesel boosting model for biodiesel yield prediction, showing (a) a waterfall plot for a representative prediction and (b) a beeswarm plot summarizing SHAP values for all input variables in the testing dataset. A full-dataset robustness check also confirmed an identical feature sequence.

significant negative SHAP values, reaching as low as −12. This aligns with the thermochemical reality of biodiesel synthesis, where excessive heat promotes methanol vaporization and catalyst inhibition, thereby reducing conversion efficiency. Reaction times also exhibit a non-monotonic trend; lower reaction times (blue points) contributed positively to the model output, while longer durations (red points) shifted into the negative SHAP region. This suggests a kinetic threshold where reversible transesterification may occur. The oil-to-MeOH ratio and catalyst concentration followed in global importance; both showed that higher values (red points) generally exerted a positive influence on the yield, though their impact magnitudes were secondary to the dominant thermal and kinetic parameters.

Overall, the SHAP analysis characterizes temperature (°C) and reaction times (min) as the dominant operational drivers within the ensemble framework. By quantifying their influence on predicted yield without proposing new reaction mechanisms, the analysis identifies these parameters as primary controllers in the alkaline transesterification process, while the oil-to-MeOH ratio and catalyst concentration exert secondary, context-dependent effects. These findings emphasize that efficient WCO conversion is governed mainly by kinetic control and the mitigation of thermal inhibition (overheating), rather than a linear increase in catalyst dosage. Moreover, the SHAP-based analysis directly contrasts with the weak or negative Pearson correlation reported in Fig. 3, demonstrating that the boosting ensemble captures non-linear and non-monotonic relationships that conventional linear correlation metrics cannot represent. However, the SHAP values delineate the computational logic of the trained ensemble. These results represent process-level sensitivities within the modeled domain and are not presented as first-principles kinetic constants.

3.6. External validation using literature-derived datasets

To rigorously assess the generalizability of the finalized AdaBoost-KNN boosting ensemble (3-fold CV), the model was benchmarked against an independent external dataset comprising 123 samples compiled from five peer-reviewed biodiesel studies. These records were strictly excluded from all phases of model development (Annexure II). To ensure architectural compatibility, a systematic unit-standardization

protocol was performed. First, catalyst concentration (wt%) was normalized to a per-unit oil basis, and oil-to-methanol molar ratios were converted to mass (g) using the molecular weights of the respective reactants, ensuring all features were defined consistently with the in-house experiments. These harmonized inputs were maintained in their original experimental scales without additional feature scaling. This approach is consistent with the direct-features architecture used during model training. Uncertainties associated with differences in reporting conventions across studies are therefore expected to be minor at the scale of the aggregated performance metrics. The optimized boosting ensemble model, configured with the hyperparameters reported in Table 1 and evaluated using 3-fold CV, achieved a high mean coefficient of determination ($R^2 = 0.973$) along with low error magnitudes (MAE = 0.874%, and RMSE = 1.729%) and their corresponding standard deviations, as illustrated in the dual-panel performance evaluation in Fig. 10(a) and detailed in Annexure IX. These results demonstrate strong external predictive performance within related WCO-based transesterification datasets, with the model explaining 97.3% of the variance in biodiesel yield across the literature-derived validation records. The predicted vs. actual yield plot (Fig. 10(b)) shows training and validation points tightly clustered along the ideal diagonal for validation folds. The narrow scatter around the diagonal and the overlapping marginal distributions along the axes suggest that the ensemble maintains high predictive accuracy and low bias without overfitting across the full yield range. This cross-study fidelity supports the model's utility as a robust biodiesel yield prediction framework for WCO transesterification.

3.7. Comparison of machine learning models with response surface methodology (RSM)

The proposed ML framework offers several advantages over traditional RSM, which is commonly used for biodiesel process optimization. The main advantage of the ensemble model is its superior ability to capture complex, non-linear kinetic interactions without being constrained by the rigid quadratic approximations typical of RSM (Dharmalingam et al., 2023; Selvan et al., 2018; Sai Bharadwaj et al., 2023). While RSM is effective for local optimization within a restricted experimental design, the boosting ensemble effectively utilized the 49-point experimental dataset and achieved high predictive accuracy ($R^2 = 0.987$). A recognized limitation of the ML approach compared with RSM is the lack of explicit parametric p-values, which RSM naturally provides for individual parameters. However, the ensemble model compensates for this limitation by providing a robust data-driven representation of non-linear variable interactions in the WCO transesterification process.

3.8. Study limitations and future outlook

This study is based on WCO-derived biodiesel production using 49 in-house experimental data points from a Tokyo restaurant and a two-step acid-base catalytic route. External validation using 123 literature-derived records from three distinct catalytic systems (homogeneous, heterogeneous, and biocatalytic) demonstrated strong predictive performance ($R^2 = 0.973$). However, absolute feedstock compositional comparability across all source studies was lacking, as parameters such as FFA content, fatty acid profile, moisture content, and AV were not consistently reported. Future work should expand the dataset to include WCO from diverse sources, geographic locations, bifunctional catalytic systems, reactor configurations, and production scales to further strengthen the framework's broader applicability. Additionally, future efforts could extend the present ensemble architecture toward a multi-objective optimization framework by incorporating key industrial evaluation indicators, such as biodiesel quality specifications, production costs, energy consumption, life cycle assessment, and carbon emissions — thereby enabling a more comprehensive and industrially relevant process design tool (Corral-Bobadilla et al., 2022; Lopresto et al., 2025).

4. Conclusions

This study developed an automated ensemble machine learning (AutoML) framework for predicting biodiesel yield from waste cooking oil (WCO), utilizing 49 in-house experimental data points from our previously validated Process-I study, while 123 literature-derived samples were used exclusively for external validation. The AutoML PyCaret environment was employed to benchmark 25 regression algorithms, with the k-nearest neighbors (KNN) model identified as the best-performing individual learner ($R^2 = 0.689$). To enhance predictive fidelity, the top 15 base learners were optimized via Optuna's tree-structured Parzen estimators (TPE) Bayesian hyperparameter tuning

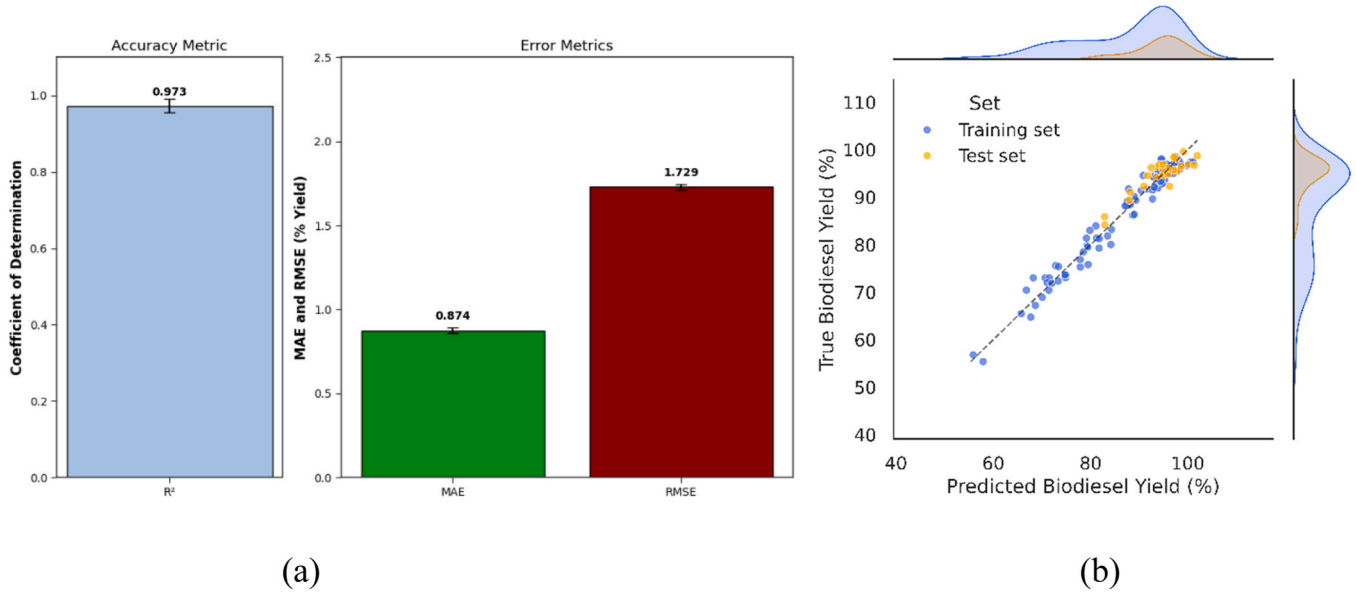

Fig. 10. External validation of the developed 3-fold boosting ensemble model using literature-derived biodiesel datasets: (a) comparative evaluation of R^2 (left panel, dimensionless) and MAE and RMSE (right panel, % yield) with error bars representing standard deviation; and (b) predicted versus actual biodiesel yield (%) for the training and validation sets.

and integrated into bagging, boosting, and stacking ensemble strategies. The AdaBoost-KNN boosting ensemble emerged as the best-performing architecture, outperforming both the bagging and stacking models. It achieved superior predictive accuracy on the experimental dataset ($R^2 = 0.987$) and demonstrated strong external predictive performance on the literature-derived dataset ($R^2 = 0.973$). Shapley Additive exPlanations (SHAP) analysis was used to decode the model's internal logic, identifying reaction temperature and time as the primary operational descriptors within the modeled process space, followed by secondary influences from the oil-to-methanol ratio and catalyst loading. Specifically, the analysis highlighted a critical 'thermal-kinetic' threshold, where the avoidance of overheating is more important for yield stability than simply increasing catalyst dosage. From a practical standpoint, the developed framework offers three immediate operational implications. First, the framework enables accurate yield prediction from limited experimental data, substantially reducing reliance on trial-and-error experimentation. Second, the SHAP-derived thermo-kinetic boundaries provide actionable guidance for defining efficient operating windows. Third, by demonstrating strong generalizability, the framework supports scalable WCO valorization and circular economy-oriented biodiesel process development across broader experimental contexts. Beyond its practical relevance, this study introduces three key innovations in WCO-based biodiesel ML research: (i) automated benchmarking of 25 regression algorithms within a unified ensemble pipeline using the PyCaret environment, minimizing manual model selection bias; (ii) integration of Optuna's TPE Bayesian optimization with SHAP interpretability, enabling transparent mapping of WCO transesterification kinetics; and (iii) rigorous external validation against 123 independently sourced records, establishing a generalizability benchmark for biodiesel ML frameworks. Overall, the proposed framework provides a robust, interpretable, and externally validated data-driven tool for WCO-based biodiesel yield prediction within the studied domain.

CRediT authorship contribution statement

Md. Rubel: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Cries Avian:** Validation, Software, Methodology, Investigation, Writing – review & editing. **M. M. Harussani:** Writing – review & editing. **Eric Kolor:** Writing – review & editing. **Sasipa Boonyubol:** Writing – review & editing, Validation, Investigation. **Koichi Mikami:** Writing – review & editing, Investigation. **Muhammad Aziz:** Writing – review & editing, Investigation. **Jeffrey S. Cross:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT Plus, Perplexity AI, QuillBot, Google Gemini and Grammarly to paraphrase and edit the language clarity while ensuring that all research findings and interpretations remained original. Additionally, after using these tools, the author(s) carefully reviewed and edited the content as required and take full responsibility for the integrity and content of the published article.

Consent to publish

All authors reviewed and approved the manuscript's content and have granted explicit permission for its publication.

Funding

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, through the Japanese Government Scholarship program awarded to the first author. There is no specific grant ID associated with this scholarship.

Nomenclature

A complete list of abbreviations and nomenclatures used in this study is provided in the supplemental material (Annexure X).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the local restaurant authority in Tokyo for their invaluable support and cooperation in facilitating sample collection for this study. The first author gratefully acknowledges the research internship conducted at the Institute of Industrial Science, The University of Tokyo, as well as the financial support provided by the Japan Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, through the Japanese Government Scholarship program for graduate studies.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.cherd.2026.05.027.

Data availability

The data supporting the key findings of this study are available in the supplementary data file. Further explanation can be provided upon request.

References

Abusweireh, R.S., Rajamohan, N., Vasseghian, Y., 2022. Enhanced production of biodiesel using nanomaterials: a detailed review on the mechanism and influencing factors. Fuel 319, 123862.
Agrawal, P., Gnanaprakash, R., Dhawane, S.H., 2024. Prediction of biodiesel yield employing machine learning: interpretability analysis via Shapley additive explanations. Fuel 359, 130516.
Agrawal, P., Rao, S., 2012. Energy-aware scheduling of distributed systems using cellular automata. IEEE, pp. 1–6.
Agrawal, P., Rao, S., 2014. Energy-aware scheduling of distributed systems. IEEE Trans. Autom. Sci. Eng. 11, 1163–1175.
Agrawal, P., Rao, S., 2021. Energy-efficient scheduling: classification, bounds, and algorithms. Sādhanā 46, 46.
Ahmad, A., Yadav, A.K., Singh, A., 2023. Application of machine learning and genetic algorithms to the prediction and optimization of biodiesel yield from waste cooking oil. Korean J. Chem. Eng. 40, 2941–2956.
Ahsan, M., Mahmud, M., Saha, P., Gupta, K., Siddique, Z., 2021. Effect of data scaling methods on machine learning algorithms and model performance. Technologies 9, 52. https://doi.org/10.3390/technologies9030052.
Ali, L., Fadhil, A., 2013. Biodiesel production from spent frying oil of fish via alkali-catalyzed transesterification. Energy Sources Part A Recovery Util. Environ. Eff. 35, 564–573.
Ali M. PyCaret: an open source, low-code machine learning library in Python. PyCaret version 1.0. 0 2020a.
Ali, M., 2020b. PyCaret: An open source, low-code machine learning library in Python. PyCaret Version 2, 514.
Almohana, A.I., Almojil, S.F., Kamal, M.A., Alali, A.F., Kamal, M., Alkhatib, S.E., et al., 2022. Theoretical investigation on optimization of biodiesel production using waste cooking oil: Machine learning modeling and experimental validation. Energy Rep. 8, 11938–11951.
Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A.B., Alzakari, N., et al., 2021. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. Appl. Sci. 11, 796.
Altikriti, E., Fadhil, A., Dheyab, M., 2015. Two-step base catalyzed transesterification of chicken fat: Optimization of parameters. Energy Sources Part A Recovery Util. Environ. Eff. 37, 1861–1866.

Asadi, B., Hajj, R., 2024. Prediction of asphalt binder elastic recovery using tree-based ensemble bagging and boosting models. Constr. Build. Mater.
Ascher, S., Watson, I., You, S., 2022. Machine learning methods for modelling the gasification and pyrolysis of biomass and waste. Renew. Sustain. Energy Rev. 155, 111902.
Avian, C., Leu, J.-S., Putro, N.A.S., Mahali, M.I., Azmi, I., Abubakar, W., et al., 2024. Temporal spatial and signal descriptor feature extraction for electronic nose signal processing. IEEE Sens. J.
Azhar, B., Taipabu, M.I., Avian, C., Viswanathan, K., Wu, W., Lau, R., 2025. Artificial intelligence-driven modeling of biodiesel production from fats, oils, and grease (FOG) with process optimization via particle swarm optimization. Energy Convers. Manag. X 26, 101000. https://doi.org/10.1016/j.ecmx.2025.101000.
Bailly, A., Blanc, C., Francis, É., Guillotin, T., Jamal, F., Wakim, B., et al., 2022. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. Comput. Methods Prog. Biomed. 213, 106504.
Baskar, G., Selvakumari, I.A.E., Aiswarya, R., 2018. Biodiesel production from castor oil using heterogeneous Ni doped ZnO nanocatalyst. Bioresour. Technol. 250, 793–798.
Bastos, R.R.C., da Luz Corrêa, A.P., da Luz, P.T.S., da Rocha Filho, G.N., Zamian, J.R., da Conceição, L.R.V., 2020. Optimization of biodiesel production using sulfonated carbon-based catalyst from an amazon agro-industrial waste. Energy Convers. Manag. 205, 112457.
Divya bhagat. Understanding Ensemble Methods: Bagging, Boosting, and Stacking. 2024.
Bhatia, S.K., Bhatia, R.K., Jeon, J.-M., Pugazhendhi, A., Awasthi, M.K., Kumar, D., et al., 2021. An overview on advancements in biobased transesterification methods for biodiesel production: oil resources, extraction, biocatalysts, and process intensification technologies. Fuel 285, 119117.
Bilanovic, D., Andargatchew, A., Kroeger, T., Shelef, G., 2009. Freshwater and marine microalgae sequestering of CO2 at different C and N concentrations–response surface methodology analysis. Energy Convers. Manag. 50, 262–267.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Buasri, A., Sirikoom, P., Pattane, S., Buachum, O., Loryuenyong, V., 2023. Process optimization of biodiesel from used cooking oil in a microwave reactor: A case of machine learning and Box–Behnken design. ChemEngineering 7, 65.
Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A., 2018. Machine learning for molecular and materials science. Nature 559, 547–555.
Cai, Z.-Z., Wang, Y., Teng, Y.-L., Chong, K.-M., Wang, J.-W., Zhang, J.-W., et al., 2015. A two-step biodiesel production process from waste cooking oil via recycling crude glycerol esterification catalyzed by alkali catalyst. Fuel Process. Technol. 137, 186–193.
Chen, Zhang, J., Luo L, B., Zhang, F., Yi, Y., Shan, Y., et al., 2021. A review on recycling techniques for bioethanol production from lignocellulosic biomass. Renew. Sustain. Energy Rev. 149, 111370.
Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Peerj Comput. Sci. 7, e623.
Corral-Bobadilla, M., Lostado-Lorza, R., Somovilla-Gómez, F., Íñiguez-Macedo, S., 2022. Life cycle assessment multi-objective optimization for eco-efficient biodiesel production using waste cooking oil. J. Clean. Prod. 359, 132113.
Degfie, T.A., Mamo, T.T., Mekonnen, Y.S., 2019. Optimized biodiesel production from waste cooking oil (WCO) using calcium oxide (CaO) nano-catalyst. Sci. Rep. 9, 18982.
Dharmalingam, B., Balamurugan, S., Wetwatana, U., Tongnan, V., Sekhar, C., Paramasivam, B., et al., 2023. Comparison of neural network and response surface methodology techniques on optimization of biodiesel production from mixed waste cooking oil using heterogeneous biocatalyst. Fuel 340, 127503.
Dhawane, S.H., Halder, G., 2019. Synthesis of catalyst support from waste biomass for impregnation of catalysts in biofuel production. Advances in Feedstock Conversion Technologies for Alternative Fuels and Bioproducts. Elsevier, pp. 199–220.
Dhawane, S.H., Kumar, T., Halder, G., 2016a. Biodiesel synthesis from Hevea brasiliensis oil employing carbon supported heterogeneous catalyst: Optimization by Taguchi method. Renew. Energy 89, 506–514.
Dhawane, S.H., Kumar, T., Halder, G., 2016b. Parametric effects and optimization on synthesis of iron (II) doped carbonaceous catalyst for the production of biodiesel. Energy Convers. Manag. 122, 310–320.
Dong, C., Wang, L., Wang, Z., Lu, J., Ding, J., 2025. A Review on Graphene-Based Catalysts for Biodiesel Production. Energy & Fuels 39, 19551–19573.
Elmaz, F., Yücel, Ö., Mutlu, A.Y., 2020. Predictive modeling of biomass gasification with machine learning-based regression methods. Energy 191, 116541.
Fadhil, A.B., Saleh, L.A., Altamer, D.H., 2020. Production of biodiesel from non-edible oil, wild mustard (Brassica juncea L.) seed oil through cleaner routes. Energy Sources Part A Recovery Util. Environ. Eff. 42, 1831–1843.
Feng, D., Guo, X., Lin, R., Xia, A., Huang, Y., Liao, Q., et al., 2021. How can ethanol enhance direct interspecies electron transfer in anaerobic digestion? Biotechnol. Adv. 52, 107812.
Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.
Girish, N., Niju, S.P., Begum, K.M.M.S., Anantharaman, N., 2013. Utilization of a cost effective solid catalyst derived from natural white bivalve clam shell for transesterification of waste frying oil. Fuel 111, 653–658.
Gonçalves, M.A., dos Santos, H.C.L., da Silva, M.A.R., da Cas Viegas, A., da Rocha Filho, G.N., da Conceição, L.R.V., 2024. Biodiesel production from waste cooking oil using an innovative magnetic solid acid catalyst based on Ni–Fe ferrite: RSM-BBD optimization approach. J. Ind. Eng. Chem. 135, 270–285.
Gouran, A., Aghel, B., Nasirmanesh, F., 2021. Biodiesel production from waste cooking oil using wheat bran ash as a sustainable biomass. Fuel 295, 120542.
Greener, J.G., Kandathil, S.M., Moffat, L., Jones, D.T., 2022. A guide to machine learning for biologists. Nat. Rev. Mol. Cell Biol. 23, 40–55.
Gupte, A.P., Basaglia, M., Casella, S., Favaro, L., 2022. Rice waste streams as a promising source of biofuels: feedstocks, biotechnologies and future perspectives. Renew. Sustain. Energy Rev. 167, 112673.
Hassan, M.M., Fadhil, A.B., 2025. Development of an effective solid base catalyst from potassium based chicken bone (K-CBs) composite for biodiesel production from a mixture of non-edible feedstocks. Energy Sources Part A Recovery Util. Environ. Eff. 47, 8056–8071.
Hong, I.K., Jeon, H., Kim, H., Lee, S.B., 2016. Preparation of waste cooking oil based biodiesel using microwave irradiation energy. J. Ind. Eng. Chem. 42, 107–112.
Huang, Y., Li, F., Bao, G., Xiao, Q., Wang, H., 2022b. Modeling the effects of biodiesel chemical composition on iodine value using novel machine learning algorithm. Fuel 316, 123348.
Huang, Z., Manzo, M., Xia, C., Cai, L., Zhang, Y., Liu, Z., et al., 2022a. Effects of waste-based pyrolysis as heating source: meta-analyze of char yield and machine learning analysis. Fuel 318, 123578.
India's WCO. Diesel doped with biodiesel made from used cooking oil rolled out. Economic Times. ⟨https://energy.economictimes.indiatimes.com/news/oil-and-gas/diesel-doped-with-biodiesel-made-from-used-cooking-oil-rolled-out/82398223⟩; 2023.
Jayaprabakar, J., Dawn, S., Ranjan, A., Priyadharsini, P., George, R., Sadaf, S., et al., 2019. Process optimization for biodiesel production from sheep skin and its performance, emission and combustion characterization in CI engine. Energy 174, 54–68.
Katongtung, T., Onsree, T., Tippayawong, N., 2022. Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes. Bioresour. Technol. 344, 126278.
Khan, H.M., Iqbal, T., Ali, C.H., Javaid, A., Cheema, I.I., 2020. Sustainable biodiesel production from waste cooking oil utilizing waste ostrich (Struthio camelus) bones derived heterogeneous catalyst. Fuel 277, 118091.
Kim, Y.-S., Kim, M.K., Fu, N., Liu, J., Wang, J., Srebric, J., 2025. Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models. Sustain. Cities Soc. 118, 105570. https://doi.org/10.1016/j.scs.2024.105570.
Kodgire, P., Sharma, A., Kachhwaha, S.S., 2023. Optimization and kinetics of biodiesel production of Ricinus communis oil and used cottonseed cooking oil employing synchronised 'ultrasound+ microwave'and heterogeneous CaO catalyst. Renew. Energy 212, 320–332.
Lin, H., Eggesbø, M., Peddada, S.D., 2022. Linear and nonlinear correlation estimators unveil undescribed taxa interactions in microbiome data. Nat. Commun. 13, 4946.
Liu, J., Zhang, Z., Tang, S., Yu, Z., Zhang, Y., Zheng, B., 2023. Ultrasound-assisted production of biodiesel from field pennycress (Thlaspi arvense L.) seeds: Process optimization and quality evaluation. Ind. Crops Prod. 203, 117224.
Lopresto, C.G., Paletta, R., Scarpello, A., Calabrò, V., 2025. Biodiesel production from waste cooking oils–Application of green chemistry principles to the multi-objective optimisation of alkaline transesterification. Chemosphere 387, 144674.
Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., et al., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.
Lyman, R., 2025. Global energy demand: A summary of the 2025 Statistical Review of World Energy. Friends Sci. Soc. ⟨https://blog.friendsofscience.org/wp-content/uploads/2025/06/STATISTICAL-REVIEW-OF-WORLD-ENERGY-2025final.pdf⟩.
Maddikeri, G.L., Pandit, A.B., Gogate, P.R., 2012. Intensification approaches for biodiesel synthesis from waste cooking oil: a review. Ind. & Eng. Chem. Res. 51, 14610–14628.
Mairizal, A.Q., Awad, S., Priadi, C.R., Hartono, D.M., Moersidik, S.S., Tazerout, M., et al., 2020. Experimental study on the effects of feedstock on the properties of biodiesel using multiple linear regressions. Renew. Energy 145, 375–381.
Mathew, G.M., Raina, D., Narisetty, V., Kumar, V., Saran, S., Pugazhendi, A., et al., 2021. Recent advances in biodiesel production: Challenges and solutions. Sci. Total Environ. 794, 148751.
Ma, T., Guo, Z., Lin, M., Wang, Q., 2021. Recent trends on nanofluid heat transfer machine learning research applied to renewable energy. Renew. Sustain. Energy Rev. 138, 110494.
Mohadesi, M., Aghel, B., Gouran, A., Razmehgir, M.H., 2022. Transesterification of waste cooking oil using Clay/CaO as a solid base catalyst. Energy 242, 122536.
Moklis, M.H., Avian, C., Shuo, C., Boonyubol, S., Cross, J.S., 2025. Machine learning-driven prediction and optimization of selective glycerol electrocatalytic reduction into propanediols. J. Electroanal. Chem. 988, 119150. https://doi.org/10.1016/j.jelechem.2025.119150.
Moradi, G., Dehghani, S., Khosravian, F., Arjmandzadeh, A., 2013. The optimized operational conditions for biodiesel production from soybean oil and application of artificial neural networks for estimation of the biodiesel yield. Renew. Energy 50, 915–920.
Mulu, E., M'Arimi, M.M., Ramkat, R.C., 2021. A review of recent developments in application of low cost natural materials in purification and upgrade of biogas. Renew. Sustain. Energy Rev. 145, 111081.
Nadim, E., Paraskar, P., Moradi, H., Hesabi, M., Qiao, Y., Murphy, E.J., et al., 2025. Kinetic and thermodynamic analysis of hemp oil epoxidation with density functional theory insights into unsaturated fatty acid epoxidation and ring-opening reactions. Chem. Eng. J. Adv. 22, 100749.

Najaf-Abadi, M.K., Ghobadian, B., Dehghani-Soufi, M., 2024. A review on application of deep eutectic solvents as green catalysts and co-solvents in biodiesel production and purification processes. Biomass-.-. Convers. Biorefinery 14, 3117–3134.
Nazloo, E.K., Moheimani, N.R., Ennaceri, H., 2023. Graphene-based catalysts for biodiesel production: Characteristics and performance. Sci. Total Environ. 859, 160000.
Nishio, M., Nishizawa, M., Sugiyama, O., Kojima, R., Yakami, M., Kuroda, T., et al., 2018. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. PloS One 13, e0195875.
Parija, P. India's dependence on imported cooking oil to continue. Al Jazeera. ⟨https://www.aljazeera.com/economy/2022/1/24/indias-dependence-on-imported-cooking-oil-to-continue⟩; 2022.
Paryanto, I., Prakoso, T., Suyono, E.A., Gozan, M., 2019. Determination of the upper limit of monoglyceride content in biodiesel for B30 implementation based on the measurement of the precipitate in a Biodiesel–Petrodiesel fuel blend (BXX). Fuel 258, 116104.
Prajapati, N., Kachhwaha, S.S., Kodgire, P., Vij, R.K., 2024. Analysis of a novel high-speed homogenizer technique for methyl ester production using waste cooking oil: Multi-objective optimization and energy analysis. Chem. Eng. Res. Des. 203, 478–491.
Raykov, T., DiStefano, C., 2025. Evaluating Change in Adjusted R-Square and R-Square Indices: A Latent Variable Method Application. Educ. Psychol. Meas. 00131644251329178.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204.
Ribeiro, M.H.D.M., dos Santos Coelho, L., 2020. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Appl. Soft Comput. 86, 105837.
Rong, G., Li, K., Su, Y., Tong, Z., Liu, X., Zhang, J., et al., 2021. Comparison of Tree-Structured Parzen Estimator Optimization in Three Typical Neural Network Models for Landslide Susceptibility Assessment. Remote Sens. 13, 4694. https://doi.org/10.3390/rs13224694.
Rubel, M., Shuo, C., Boonyubol, S., Harussani, M., Kachhwaha, S.S., Cross, J.S., 2026. Optimization of a Two-Step Biodiesel Production from Waste Cooking Oil: Comparative Evaluation of n-Hexane and CPME as Transesterification Cosolvents. Chem. Eng. Res. Des.
Sai Bharadwaj, A., S, N., Begum, K.M.S., N, A., 2023. Free fatty acid optimization and modeling of biodiesel production from high viscous rubber seed oil–A comparative study of RSM and ANN. Energy Sources Part A Recovery Util. Environ. Eff. 45, 3475–3489.
Schober, P., Boer, C., Schwarte, L.A., 2018. Correlation coefficients: appropriate use and interpretation. Anesth. & Analg. 126, 1763–1768.
Schweidtmann, A.M., Esche, E., Fischer, A., Kloft, M., Repke, J., Sager, S., et al., 2021. Machine learning in chemical engineering: A perspective. Chem. Ing. Tech. 93, 2029–2039.
Sebayang, A.H., Ideris, F., Silitonga, A.S., Shamsuddin, A., Zamri, M., Pulungan, M.A., et al., 2023b. Optimization of ultrasound-assisted oil extraction from Carica candamarcensis; A potential Oleaginous tropical seed oil for biodiesel production. Renew. Energy 211, 434–444.
Sebayang, A., Kusumo, F., Milano, J., Shamsuddin, A., Silitonga, A., Ideris, F., et al., 2023a. Optimization of biodiesel production from rice bran oil by ultrasound and infrared radiation using ANN-GWO. Fuel 346, 128404.
Selvan, S.S., Pandian, P.S., Subathira, A., Saravanan, S., 2018. Comparison of response surface methodology (RSM) and artificial neural network (ANN) in optimization of Aegle marmelos oil extraction for biodiesel production. Arab. J. Sci. Eng. 43, 6119–6131.
Sieradzki S., Mańdziuk J. Modified Adaptive Tree-Structured Parzen Estimator for Hyperparameter Optimization. arXiv Preprint arXiv:250200871 2025.
Silitonga, A.S., Shamsuddin, A.H., Mahlia, T.M.I., Milano, J., Kusumo, F., Siswantoro, J., et al., 2020. Biodiesel synthesis from Ceiba pentandra oil by microwave irradiation-assisted transesterification: ELM modeling and optimization. Renew. Energy 146, 1278–1291.
Snoek, J., Larochelle, H., Adams, R.P., 2012. Practical Bayesian Optimization of Machine Learning Algorithms. In: Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems, 25. Curran Associates, Inc.
Stamenković, O.S., Rajković, K., Veličković, A.V., Milić, P.S., Veljković, V.B., 2013. Optimization of base-catalyzed ethanolysis of sunflower oil by regression and artificial neural network models. Fuel Process. Technol. 114, 101–108.
Statista Inc (2023). Food self-sufficiency ratio of oils and fats in Japan. Statista. ⟨https://www.statista.com/statistics/1040066/japan-food-self-sufficiency-ratio-oils-fats/⟩. Statista Inc.; 2023.
Sultana, N., Hossain, S.Z., Abusaad, M., Alanbar, N., Senan, Y., Razzak, S., 2022. Prediction of biodiesel production from microalgal oil using Bayesian optimization algorithm-based machine learning approaches. Fuel 309, 122184.
Tang, Z.-C., Zhenzhou, L., Zhiwen, L., Ningcong, X., 2015. Uncertainty analysis and global sensitivity analysis of techno-economic assessments for biodiesel production. Bioresour. Technol. 175, 502–508.
Tao S., Peng P., Li Q., Wang H. Supervised Contrastive Learning with Tree-Structured Parzen Estimator Bayesian Optimization for Imbalanced Tabular Data. arXiv Preprint arXiv:221010824 2022.
The Asahi Shimbun Company. (2024, July 30). Japan's food self-sufficiency ratio hits record low. Asahi Shimbun. ⟨https://www.asahi.com/ajw/articles/14704414⟩. The Asahi Shimbun Company; 2024.
Thushari, I., Babel, S., 2018. Sustainable utilization of waste palm oil and sulfonated carbon catalyst derived from coconut meal residue for biodiesel production. Bioresour. Technol. 248, 199–203.
Toyao, T., Maeno, Z., Takakusagi, S., Kamachi, T., Takigawa, I., Shimizu, K., 2019. Machine learning for catalysis informatics: recent applications and prospects. Acs Catal. 10, 2260–2297.
Usman, M., Cheng, S., Cross, J.S., 2023. Biodiesel production from wet sewage sludge and reduced CO2 emissions compared to incineration in Tokyo, Japan. Fuel 341, 127614.
Wang, L., Dong, C., Yuan, Z., Wang, Z., Lu, J., Ding, J., 2025. Z-type ZnO/CeO2@ g-CN photocatalyst for efficient pre-esterification of waste cooking oil: Optimization, kinetics, and thermodynamics. Renew. Energy, 124919.
Wang, J., Wang, S., 2019. Preparation, modification and environmental application of biochar: A review. J. Clean. Prod. 227, 1002–1022.
Watanabe S. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. arXiv Preprint arXiv:230411127 2023.
Wolpert, D.H., 1992. Stacked generalization. Neural Netw. 5, 241–259.
Yuan, Z., Zhu, J., Dong, C., Wang, L., Lu, J., Li, Y., et al., 2024a. RSM optimization and kinetics study of α-Fe2O3/g-C3N4@ H photocatalyst with S-type heterojunction for esterifying crude castor oil to reduce acidity and synthesize biodiesel. Energy Convers. Manag. 309, 118449.
Yuan, Z., Zhu, J., Lu, J., Li, Y., Ding, J., 2024b. Preparation of biodiesel by transesterification of castor oil catalyzed by flaky halloysite supported ZnO/SnO2 heterojunction photocatalyst. Renew. Energy 227, 120516.
Zakir Hossain, S., Sultana, N., Irfan, M.F., Haque, S.M., Nasr, N., Razzak, S.A., 2022. Artificial intelligence-based super learner approach for prediction and optimization of biodiesel synthesis—A case of waste utilization. Int. J. Energy Res. 46, 20519–20534.
Zhang, X., Li, H., Sekar, M., Elgendi, M., Krishnamoorthy, N., Xia, C., et al., 2023a. Machine learning algorithms for a diesel engine fuelled with biodiesel blends and hydrogen using LSTM networks. Fuel 333, 126292.
Zhang, L., Xing, X., Liu, Y., Shi, W., Wang, M., 2023b. Directional methanolysis of kitchen waste for the co-production of methyl levulinate and fatty acid methyl esters: Catalytic strategy and machine learning modeling. Bioresour. Technol. 367, 128274.
Zhou, H., Huang, W., Xiao, Z., Zhang, S., Li, W., Hu, J., et al., 2022. Deep-learning-assisted noncontact gesture-recognition system for touchless human-machine interfaces. Adv. Funct. Mater. 32, 2208271.
Zhu, J., Yuan, Z., Wang, L., Dong, C., Lu, J., Ding, J., 2024. 3D simulation in a fixed bed coupled pervaporation reactor for biodiesel production. Chem. Eng. J. 496, 153854.