Spectrum of Engineering Sciences
ISSN (e) 3007-3138 **(p)** 3007-312X
Volume 4, Issue 3, 2026

EXPLAINABLE MACHINE LEARNING FRAMEWORK FOR CATALYTIC PYROLYSIS PARAMETER OPTIMIZATION IN SUSTAINABLE HYDROGEN PRODUCTION

Subhan Azeem[*1], **Nadeem Hassan**[2], **Muhammad Ashraf**[3]

[*1,2]*NFC Institute of Engineering & Technology, Multan*
[3]*Bahauddin Zakriya University, Multan*

[*1]msazeem@nfciet.edu.pk,[2]nadeemhassan@nfciet.edu.pk,[3]ashraf_ce@bzu.edu.pk

DOI: https://doi.org/10.5281/zenodo.19626750

Keywords
Explainable machine learning, Catalytic pyrolysis, Sustainable hydrogen production, Biomass conversion, SHAP analysis, Process optimization

Article History
Received: 01 February 2026
Accepted: 05 March 2026
Published: 14 March 2026

Corresponding Author: *
Subhan Azeem

Abstract

This study proposes an explainable machine learning framework for optimizing catalytic pyrolysis parameters for sustainable hydrogen production from biomass. The framework is designed to model the highly nonlinear relationship between feedstock composition, catalyst properties, and operating conditions that collectively govern hydrogen yield in catalytic pyrolysis systems. A structured experimental dataset is constructed from reported catalytic pyrolysis studies and includes biomass-related variables, catalyst descriptors, and reaction parameters as model inputs, with hydrogen yield as the continuous prediction target. Multiple supervised regression algorithms are evaluated to identify the most suitable predictive model, and comparative analysis shows that ensemble-based methods outperform conventional baseline models in capturing the multivariable and interaction-driven nature of the process. The selected model is further examined using SHAP-based explainable artificial intelligence to quantify the contribution of individual features and to reveal the dominant influence of reaction temperature, nickel loading, catalyst support, and biomass hydrogen content on hydrogen-production behavior. To move beyond forward prediction, the trained model is integrated into an optimization framework that explores the process-variable space and identifies parameter combinations associated with improved hydrogen yield under practically meaningful operating bounds. The results demonstrate strong agreement between predicted and observed hydrogen-yield values, stable residual behavior, and a well-defined high-performance operating region in the optimization landscape. The explainability analysis confirms that the learned variable importance is physically consistent with known catalytic and thermochemical mechanisms, thereby strengthening the scientific credibility of the framework. Overall, the proposed approach provides a predictive, interpretable, and optimization-oriented methodology for catalytic pyrolysis analysis, offering a practical pathway to reduce experimental trial-and-error, improve process understanding, and support the rational design of more efficient and sustainable biomass-based hydrogen-production systems.

INTRODUCTION

The global transition toward low-carbon energy systems has intensified interest in hydrogen as a versatile energy carrier capable of supporting decarbonization across transport, industry, power generation, and chemical manufacturing [1]. However, the sustainability of hydrogen depends not only on its end-use profile but also on the carbon intensity, resource efficiency, and scalability of the production route. Conventional hydrogen production remains dominated by fossil-based pathways, which, although mature and economically competitive, are associated with substantial greenhouse gas emissions and continued dependence on non-renewable feedstocks [2]. In this context, biomass-derived hydrogen has emerged as an attractive alternative because lignocellulosic biomass is abundant, renewable, geographically distributed, and compatible with circular-economy principles. Among thermochemical conversion routes, catalytic pyrolysis has gained increasing attention because it can convert complex solid biomass into hydrogen-rich gas streams while simultaneously enabling valorization of bio-oil and char co-products. The appeal of catalytic pyrolysis lies in its ability to integrate thermal decomposition with catalytic upgrading in a single or closely coupled processing framework, thereby enhancing gas quality, promoting reforming and cracking reactions, and suppressing undesirable oxygenated intermediates [3]. Yet despite its promise, catalytic pyrolysis remains a highly nonlinear and multivariable process in which hydrogen yield is shaped by intertwined effects of feedstock composition, ash content, volatile matter, catalyst chemistry, metal loading, support structure, residence time, and operating temperature. These interactions complicate process design and make conventional one-factor-at-a-time experimentation inefficient for discovering robust operating windows. As a result, recent research increasingly frames catalytic pyrolysis not simply as a reaction-engineering problem, but as a complex optimization problem requiring integrated data-driven, mechanistic, and interpretable decision-support tools.

From a thermochemical standpoint, catalytic pyrolysis is especially challenging because hydrogen formation does not arise from a single dominant reaction but from a network of primary and secondary transformations occurring across multiple temporal and spatial scales [4]. Biomass first undergoes depolymerization and devolatilization, generating char, condensable vapors, and permanent gases. These primary products then experience secondary reactions such as catalytic cracking, steam reforming, dry reforming, methane reforming, deoxygenation, and water-gas shift conversions, all of which compete or cooperate depending on process conditions. Even small changes in reactor temperature or catalyst formulation can alter the relative rates of these pathways, leading to major shifts in hydrogen selectivity, tar evolution, coke deposition, and catalyst deactivation. Nickel-based catalysts remain widely studied because of their favorable cost-activity balance and their established ability to promote C–C bond cleavage and reforming reactions. Still, their performance is strongly affected by support type, metal dispersion, calcination conditions, and resistance to sintering and carbon formation [5]. Likewise, catalyst supports such as alumina, silica, dolomite, calcium-based materials, zeolitic structures, and carbonaceous supports can influence acidity-basicity, oxygen mobility, metal-support interaction, and diffusion behavior, thereby reshaping product distributions in nontrivial ways. Recent literature also emphasizes that biomass variability itself is a major source of uncertainty. Feedstocks differ widely in cellulose, hemicellulose, lignin, moisture, fixed carbon, and elemental composition, which means that a parameter combination optimized for one feedstock may not generalize well to another. Consequently, the design objective is no longer only to maximize hydrogen yield under isolated laboratory conditions, but to identify parameter regimes that remain effective across heterogeneous data and realistic operating variability. This broader optimization perspective has created fertile ground for machine learning, especially when trained on curated experimental datasets that encode relationships too intricate to capture fully using simplified empirical correlations alone [6].

Recent advances in machine learning for biomass pyrolysis and related thermochemical systems suggest that data-driven models can substantially improve predictive performance, accelerate screening of candidate conditions, and reveal latent patterns hidden within heterogeneous experimental literature. Studies published over the last two years show growing use of tree-based ensembles, gradient boosting methods, support vector machines, neural networks, and hybrid optimization frameworks to predict gas yield, bio-oil composition, biochar properties, activation energy, and hydrogen-rich syngas behavior from practical feedstock and operating descriptors [7]. These developments are important because catalytic pyrolysis datasets are typically modest in size, nonlinear in behavior, and heterogeneous in measurement conditions, making ensemble learning methods especially attractive due to their robustness, flexibility, and tolerance of mixed feature interactions. Several recent investigations have shown that machine learning models can outperform traditional regression approaches in biomass pyrolysis prediction tasks while also supporting process optimization through evolutionary algorithms or Bayesian strategies. At the same time, the literature indicates a methodological shift from purely predictive models toward more decision-oriented frameworks that combine prediction with sensitivity analysis, uncertainty-aware optimization, and inverse design. Inverse modeling is particularly relevant for hydrogen production because practitioners often begin with a target performance requirement, such as a desired hydrogen yield or gas composition, and then seek the process conditions most likely to realize that target. This reverses the conventional forward problem and elevates machine learning from a passive forecasting tool to an active process-design aid [8]. For sustainable hydrogen production, such a capability is highly valuable because it can reduce experimental burden, narrow the search space for high-performing catalyst-feedstock combinations, and support more rational scaling of biomass conversion technologies. Nevertheless, predictive accuracy alone is not sufficient for research or deployment, because black-box optimization can obscure the physicochemical basis of recommended conditions and limit trust in the resulting design choices.

This need for trust, transparency, and scientific interpretability has driven the increasing adoption of explainable machine learning in pyrolysis research. Explainable methods are especially valuable in catalytic pyrolysis because the process is governed by domain-relevant variables whose effects are meaningful to chemists and process engineers, such as temperature, catalyst loading, heating rate, residence time, elemental composition, and support chemistry [9]. Rather than treating model outputs as opaque predictions, explainable machine learning seeks to quantify the contribution of each input variable, uncover nonlinear threshold behavior, and expose interaction effects that may correspond to underlying reaction mechanisms. Recent studies using SHAP-based analysis and related interpretability tools have shown that process temperature repeatedly emerges as one of the strongest determinants of pyrolysis performance, while feedstock composition and catalyst-related descriptors also exert substantial influence on product distribution and hydrogen-rich gas formation. This is consistent with chemical intuition: temperature controls devolatilization severity and reforming intensity, while catalyst properties regulate cracking, reforming, and deoxygenation selectivity [10]. Explainability therefore, serves two purposes simultaneously. First, it improves confidence in model recommendations by demonstrating alignment between learned patterns and known thermochemical principles. Second, it enables scientific discovery by highlighting previously underappreciated interactions or parameter regions worthy of targeted experimentation. In the context of sustainable hydrogen production, this is especially significant because optimization should not aim merely at numerical performance maxima; it should also guide the design of catalytically efficient, operationally stable, and practically interpretable systems [11]. Explainable machine learning thus bridges the gap between predictive analytics and engineering insight, making it particularly well suited for domains where data remains limited, experimental campaigns are costly, and mechanistic plausibility matters as much as statistical fit.

Despite these encouraging developments, significant research gaps remain at the intersection of catalytic pyrolysis, sustainable hydrogen production, and explainable data-driven optimization. A large portion of existing studies in biomass thermochemical conversion focus on overall product yields such as bio-oil, char, or gas, rather than hydrogen-specific optimization under catalytically assisted conditions. Even when hydrogen is considered, many studies emphasize forward prediction only, with limited treatment of inverse parameter selection, target-driven optimization, or transparent ranking of the most influential factors. Moreover, the available literature often addresses either catalyst design, feedstock effects, or operating-condition tuning in isolation, whereas practical hydrogen optimization requires a unified view of these dimensions. Another limitation is that many models are developed for narrow datasets, single feedstock families, or simplified feature sets, which restricts generalizability and weakens their utility for broader process planning. In addition, recent reviews have noted that the role of explainability in biomass pyrolysis is still emerging, and that integrating XAI with optimization remains a particularly promising but underdeveloped direction. For catalytic pyrolysis, this gap is consequential: researchers need frameworks that not only predict hydrogen yield accurately but also identify why particular conditions are effective, how sensitive those recommendations are to perturbations, and whether the suggested operating windows align with known catalytic and thermochemical behavior. A further challenge concerns the tension between process performance and sustainability objectives. Maximizing hydrogen alone may lead to conditions that increase energy demand, accelerate catalyst deactivation, or reduce practical viability. Therefore, the concept of "sustainable hydrogen production" requires broader optimization thinking that balances yield enhancement with catalyst efficiency, feedstock flexibility, and experimentally plausible operating regimes. These unresolved issues motivate the development of a more transparent, integrated, and optimization-oriented framework for catalytic pyrolysis analysis [12].

Against this background, the present study proposes an explainable machine learning framework for optimizing catalytic pyrolysis parameters for sustainable hydrogen production. The central premise is that recent progress in data-driven modeling, coupled with interpretable analytics, now makes it possible to move beyond descriptive experimentation toward rational parameter discovery grounded in both predictive performance and scientific insight. By learning from curated catalytic pyrolysis data and systematically relating hydrogen yield to feedstock descriptors, catalyst attributes, and process conditions, the framework is intended to identify the combinations most conducive to enhanced hydrogen generation while also clarifying the relative importance of the governing factors. The explainable component is essential because it transforms the model from a black-box estimator into an interpretable support tool capable of revealing whether hydrogen optimization is driven primarily by thermal severity, catalyst composition, support effects, biomass hydrogen content, or interaction among these variables. In methodological terms, such a framework supports both forward understanding and inverse decision-making: it can estimate hydrogen outcomes under specified conditions and also guide the search for parameter settings likely to achieve improved performance. In scientific terms, it offers a structured means of connecting statistical learning with catalytic pyrolysis knowledge [13]. In practical terms, it has the potential to reduce costly trial-and-error experimentation, improve parameter screening efficiency, and support the design of more reliable hydrogen-oriented biomass conversion strategies. Accordingly, this study is positioned within the latest literature not merely as another predictive modeling exercise, but as a step toward interpretable, optimization-driven, and sustainability-aware process intelligence for biomass-based hydrogen production.

Literature review

Catalytic pyrolysis has become an important thermochemical route for converting lignocellulosic biomass into hydrogen-rich gaseous products because it combines biomass decomposition with catalytic upgrading in a relatively compact processing framework. Recent literature shows that hydrogen formation in catalytic pyrolysis is governed by a coupled network of devolatilization, cracking, reforming, and water-gas-shift reactions rather than by a single dominant pathway, which makes process behavior highly nonlinear and sensitive to operating conditions [14]. Temperature is repeatedly identified as a principal control variable because it strongly affects tar cracking, methane reforming, gas composition, and the balance between hydrogen generation and undesired side reactions such as coke formation. Recent thermodynamic and experimental studies also report that higher-temperature regimes generally favor cracking and reforming, while lower-temperature operation is more vulnerable to incomplete conversion and carbon deposition.

A second major theme in the recent literature is the central role of catalyst design. Nickel-based catalysts remain among the most widely investigated options for catalytic pyrolysis and related reforming systems because they offer a favorable cost–activity balance and strong activity for C–C bond cleavage and reforming reactions. However, their performance is closely linked to support chemistry, calcination conditions, metal dispersion, and resistance to sintering and coking. A study on machine-learning-guided optimization of nickel-based catalysts for biohydrogen production emphasized that support type, nickel loading, and reaction temperature jointly shape hydrogen yield, and that these factors interact in ways that are difficult to optimize through conventional experimentation alone [15]. More broadly, recent hydrogen and catalysis literature has continued to show that supports such as alumina, silica-based materials, dolomite, and carbonaceous structures can substantially influence catalytic activity, product selectivity, and stability, reinforcing the view that catalyst optimization is inseparable from process optimization.

Another recurring issue is feedstock heterogeneity. Biomass feedstocks differ widely in cellulose, hemicellulose, lignin, ash, volatile matter, moisture, and elemental composition, and these differences often alter pyrolysis behavior even under nominally similar reactor conditions. Recent studies on biomass pyrolysis modeling note that data variability across feedstocks is one of the main reasons simple empirical or single-factor approaches often fail to generalize well. This has shifted the field toward multivariable and data-driven analysis, especially for hydrogen-oriented applications in which researchers seek not just acceptable gas yield, but operating conditions that remain effective across heterogeneous biomass sources. Current review work on artificial intelligence in biomass pyrolysis similarly argues that future progress depends on models that can handle diverse datasets, nonlinear interactions, and process uncertainty while remaining interpretable and transferable to real operating scenarios [11].

Over the past few years, machine learning has become increasingly prominent in biomass pyrolysis research because it can learn nonlinear relationships between feedstock descriptors, catalyst variables, and operating conditions without requiring an explicitly derived mechanistic equation for every interaction. Recent publications have applied random forests, gradient boosting models, support vector machines, neural networks, CatBoost, XGBoost, and hybrid tree-based methods to predict product yields, syngas composition, activation energy, and catalyst-performance trends in biomass pyrolysis and related conversion systems [16]. These studies consistently report that ensemble and boosting methods perform especially well on pyrolysis datasets because they are robust to moderate data size, can accommodate mixed variables, and capture strong nonlinear interactions.

A notable recent direction is the movement from pure prediction toward optimization-oriented modeling. Instead of using machine learning only to estimate pyrolysis outputs from known inputs, newer studies combine prediction models with Bayesian optimization, evolutionary search, or other inverse-design strategies to identify operating conditions likely to maximize target performance. This is especially relevant in catalytic pyrolysis for hydrogen production because the practical problem is usually target-driven: researchers often want to determine what temperature, catalyst properties, or feedstock characteristics are most likely to produce a desired hydrogen yield. Recent work shows growing interest in such inverse and optimization-guided frameworks across pyrolysis, syngas, and hydrogen-related systems. For example, Bayesian-optimization-assisted models have been used to improve biomass pyrolysis prediction and optimization, while a catalytic pyrolysis study explicitly employed a two-stage machine learning framework with Bayesian optimization to identify conditions favorable for hydrogen production [17].

The literature also indicates that machine learning is being used across adjacent hydrogen-conversion domains, not only in catalytic biomass pyrolysis. A review on machine learning applications in fuel reforming highlighted the wider adoption of ML methods in hydrogen-production studies and explicitly discussed biomass catalytic pyrolysis as an area where data-driven models can accelerate process and catalyst optimization [18]. Similarly, recent work on syngas-yield prediction from biomass pyrolysis and gasification demonstrates that tree-based and hybrid ML frameworks are increasingly viewed as practical tools for mapping complex process landscapes relevant to hydrogen-rich gas production. This broader adoption matters because it suggests that the field is gradually moving toward unified digital workflows in which prediction, optimization, and interpretation are treated as integrated tasks rather than separate modeling exercises.

Although predictive performance remains important, recent literature increasingly argues that black-box accuracy alone is not sufficient for thermochemical process research. In catalytic pyrolysis, process recommendations must be scientifically credible because model outputs are expected to align with reaction chemistry, catalyst behavior, and engineering feasibility. This has led to greater interest in explainable machine learning, particularly SHAP-based and feature-attribution approaches that help quantify how individual variables contribute to predicted outcomes. A 2025 study on explainable AI for predictive biomass pyrolysis explicitly used interpretable ML to analyze the drivers of bio-oil, biochar, and biogas yields, while broader review work on AI in biomass pyrolysis identified interpretability as a critical requirement for future deployment and scale-up [19].

Recent explainable studies suggest that temperature often emerges as one of the strongest predictors of pyrolysis outcomes, but catalyst and feedstock descriptors can become equally important when the task shifts from general product prediction to hydrogen-oriented optimization. The nickel-catalyst optimization study and the hydrogen-production study both emphasized the importance of temperature and nickel-related variables, while the latter used SHAP analysis to show that final temperature and nickel loading were among the most influential parameters in classification and regression stages. Related explainable work on activation-energy prediction in combustion and pyrolysis also confirms that interpretable ML can reveal physically meaningful structure in complex thermal-conversion datasets. Collectively, these studies support the argument that explainability is not merely an optional add-on but a scientifically useful mechanism for linking statistical predictions to process understanding [20].

Interpretability is particularly important in sustainable hydrogen production because the best numerical prediction may not always correspond to the most practically useful operating window. High hydrogen yields achieved under narrowly tuned or catalyst-intensive conditions may be less attractive if they impose excessive energy demand, rapid deactivation, or poor scalability. Explainable machine learning can partly address this issue by revealing whether model recommendations are driven by stable trends or by narrow regions in the data. Recent reviews and special-topic discussions in catalysis and pyrolysis increasingly call for interpretable, generalizable, and mechanism-aware AI systems rather than opaque models optimized only for benchmark accuracy. This trend directly supports the development of explainable frameworks for catalytic pyrolysis parameter optimization.

Despite growing progress, the literature still shows several important gaps. First, much of the machine learning work in biomass pyrolysis has focused on overall product yields such as bio-oil, gas, or char rather than on hydrogen-specific optimization under catalytic conditions. Even when hydrogen is considered, many studies emphasize forward prediction rather than inverse parameter discovery or target-driven optimization. Recent hydrogen-oriented papers have begun to address this, but the number of such studies remains limited compared with the much larger body of work on general pyrolysis yield prediction.

Second, many existing studies examine either feedstock effects, catalyst effects, or operating conditions in partial isolation. However, catalytic pyrolysis for sustainable hydrogen production is inherently multivariable, and recent evidence shows that performance emerges from interactions among biomass composition, catalyst chemistry, and thermal conditions rather than from any single factor alone. This means that a more useful framework should unify these variable groups within one predictive and optimization pipeline. Recent catalyst-optimization and syngas-prediction studies support this view by showing that multifeatured models outperform simpler approaches and can uncover hidden cross-effects that are difficult to detect manually.

Third, explainability remains underdeveloped in hydrogen-focused catalytic pyrolysis research. While recent studies increasingly use SHAP or similar tools, the literature still lacks enough frameworks that combine three elements at once: hydrogen-oriented prediction, inverse optimization, and interpretable factor analysis. The 2026 inverse-analysis study is an important step in this direction, but the field still needs broader, more systematically framed approaches that explicitly position explainability as a central part of sustainable process design rather than as a post hoc visualization step. Recent reviews on AI in pyrolysis make this limitation clear by identifying interpretability, generalizability, and integration with optimization as key future needs.

In view of these gaps, the present topic is well motivated. An explainable machine learning framework for optimizing catalytic pyrolysis parameters for sustainable hydrogen production responds directly to the latest direction of the field by integrating multivariable prediction, parameter optimization, and interpretable analysis. Such a framework is justified by three consistent findings from the recent literature: catalytic pyrolysis is strongly nonlinear and difficult to optimize experimentally; machine learning is increasingly effective for modeling pyrolysis and hydrogen-related thermochemical systems; and explainability is now seen as essential for scientific credibility, engineering trust, and sustainable deployment.

Methodology

This study proposes an explainable machine learning framework for optimizing catalytic pyrolysis parameters for sustainable hydrogen production. The overall methodology is designed to model the complex nonlinear relationship between biomass characteristics, catalyst properties, operating conditions, and hydrogen yield, while also providing interpretable insights into the contribution of each parameter. The framework integrates data curation, preprocessing, predictive modeling, model evaluation, explainability analysis, and parameter optimization within a unified workflow. Rather than relying solely on conventional trial-and-error experimentation, the proposed approach uses data-driven learning to identify high-performing operating conditions and to support more efficient process design.

The methodological pipeline consists of five major stages. In the first stage, experimental data on catalytic pyrolysis and hydrogen production are collected and organized into a structured dataset. In the second stage, the dataset is preprocessed through cleaning, encoding, normalization, and feature preparation. In the third stage, multiple machine learning models are developed and compared to predict hydrogen yield from the selected process variables. In the fourth stage, explainable artificial intelligence techniques are applied to identify the most influential variables and to interpret model behavior. In the fifth stage, the best-performing model is used in an optimization setting to determine the parameter combinations most favorable for sustainable hydrogen production. This integrated design ensures that the framework is not only predictive but also interpretable and decision-oriented.

The dataset for this study is constructed from experimental catalytic pyrolysis studies reported in the literature. Only studies reporting measurable hydrogen yield along with relevant feedstock, catalyst, and process information are included. Data points are extracted from published experimental results and compiled into a tabular dataset in which each row represents one catalytic pyrolysis condition and its corresponding hydrogen output. The purpose of this dataset construction process is to capture the multidimensional experimental variability associated with biomass-based hydrogen production.

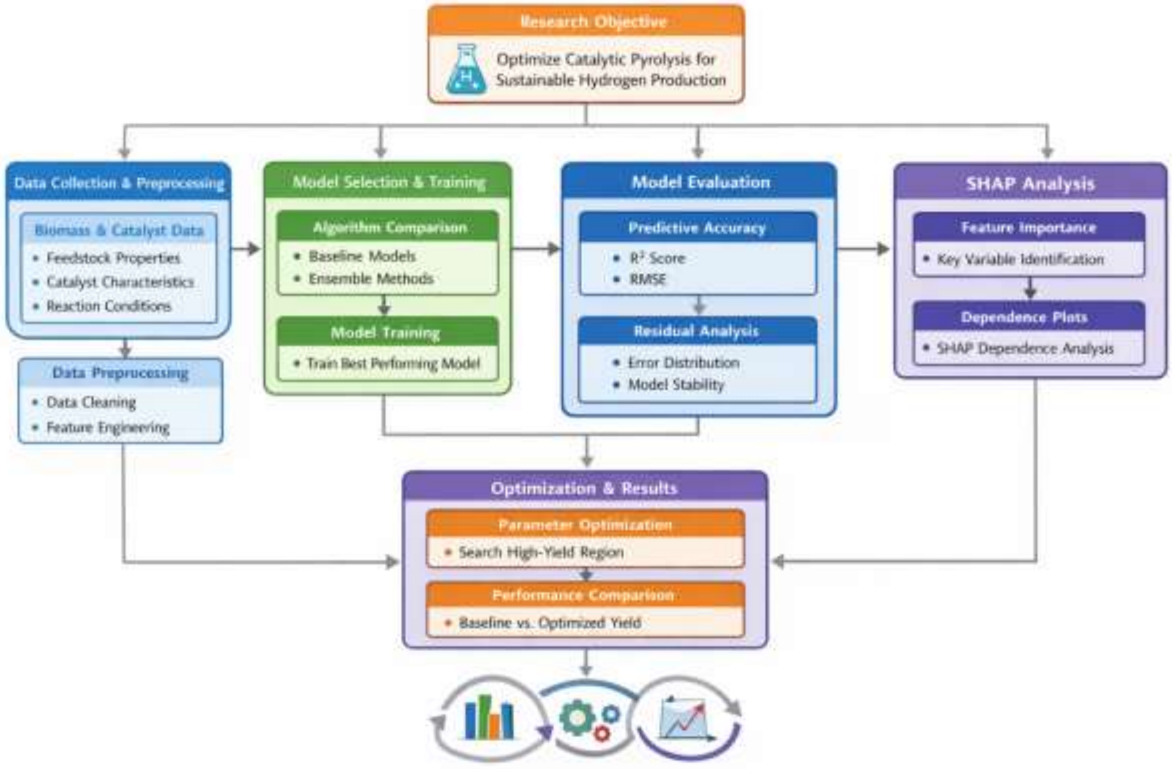

Figure 1: Methodology Flow Diagram

The input variables are grouped into three broad categories. The first category includes biomass-related properties, such as moisture content, volatile matter, ash content, fixed carbon, and elemental composition including carbon, hydrogen, oxygen, and nitrogen where available. These variables represent the physicochemical characteristics of the feedstock and are important because biomass composition strongly affects devolatilization, cracking, and reforming behavior. The second category includes catalyst-related parameters, such as catalyst type, support material, metal loading, and calcination temperature. These variables reflect the catalytic environment that governs reforming efficiency, gas selectivity, and coke resistance. The third category includes operating conditions, such as reaction temperature, heating rate, residence time, and other experimentally reported process parameters. The output variable is hydrogen yield, expressed in a consistent numerical form across the dataset. To improve data consistency, studies with missing critical variables, ambiguous measurements, or incompatible reporting units are excluded. Where necessary, numerical values are converted to common units before analysis. This structured dataset serves as the foundation for the development of machine learning models capable of learning the relationship between catalytic pyrolysis conditions and hydrogen production.

Because the dataset is compiled from heterogeneous experimental sources, preprocessing is necessary to ensure reliability and compatibility with machine learning models. The preprocessing stage begins with data cleaning, which includes removal of duplicate entries, treatment of incomplete records, correction of obvious formatting inconsistencies, and verification of variable ranges. If limited missing values are present in non-critical features, appropriate imputation techniques may be applied; however, entries missing the target variable or multiple essential process descriptors are removed to preserve dataset quality.

Categorical variables, such as catalyst support type or catalyst family, are transformed into machine-readable form using encoding techniques. One-hot encoding is employed for nominal categorical features because it preserves category identity without imposing ordinal meaning. Numerical variables are scaled using normalization or standardization methods depending on the model requirements. This step ensures that variables measured on different scales, such as temperature and metal loading, contribute more consistently during training. Feature scaling is particularly important for distance-based and gradient-based models, while tree-based models are generally less sensitive to scale; however, a consistent preprocessing pipeline is maintained across model development for methodological rigor.

Following preprocessing, the dataset is divided into input features and target output. The final feature matrix contains biomass, catalyst, and operating-condition variables, while the target vector contains hydrogen yield values. The processed dataset is then split into training and testing subsets. A common strategy is to use 80% of the data for training and 20% for testing, although this ratio may be adjusted depending on dataset size. To improve robustness and reduce the influence of random partitioning, cross-validation is also used during model training and hyperparameter tuning. This preprocessing pipeline provides a clean and structured foundation for reliable predictive modeling.

The predictive stage of the framework involves the development and comparison of several supervised machine learning models for hydrogen-yield prediction. Since hydrogen production in catalytic pyrolysis is influenced by highly nonlinear interactions, the study considers models capable of capturing both simple and complex relationships. A set of baseline and advanced regression models is evaluated, including linear regression, decision tree regression, random forest regression, gradient boosting regression, support vector regression, extreme gradient boosting, and other suitable ensemble methods. These models are selected because they collectively represent a broad spectrum of learning behavior, from simple interpretable baselines to high-capacity nonlinear predictors.

Each model is trained on the preprocessed training data using the same input feature space. Hyperparameter tuning is performed to improve predictive performance and reduce overfitting. Depending on the model, tuning may involve adjustment of tree depth, number of estimators, learning rate, regularization strength, kernel parameters, or other model-specific settings. Grid search, random search, or Bayesian optimization may be used for this purpose, with cross-validation applied to identify the parameter combination that gives the best generalization performance on unseen folds.

The trained models are then compared using standard regression evaluation metrics. These include the coefficient of determination, mean absolute error, root mean squared error, and mean absolute percentage error where suitable. The coefficient of determination measures how well the model explains the variance in hydrogen yield, while mean absolute error and root mean squared error quantify prediction error in practical terms. The model achieving the best balance between predictive accuracy, generalization, and stability is selected as the final predictive model. This selected model is then passed to the explainability and optimization stages of the framework.

Although high predictive performance is important, this study places equal emphasis on interpretability. For this reason, explainable machine learning methods are applied to the best-performing model to understand how the input variables influence hydrogen-yield prediction. Explainability analysis is essential because catalytic pyrolysis is a physically meaningful process, and model outputs must be interpretable in terms of feedstock properties, catalyst effects, and operating conditions. To achieve this, SHAP is used as the primary explainability technique. SHAP assigns each input feature a contribution value for a given prediction, allowing both global and local interpretation of model behavior. Global SHAP analysis is used to rank the overall importance of variables across the entire dataset. This helps identify which factors, such as reaction temperature, nickel loading, support material, or biomass hydrogen content, have the greatest influence on hydrogen production. Local SHAP analysis is used to interpret individual predictions and to understand why a specific parameter combination leads to a high or low predicted hydrogen yield.

In addition to feature-importance ranking, partial dependence analysis and feature-interaction examination may also be used to better understand nonlinear effects and pairwise dependencies. These tools help reveal whether a feature has a monotonic, threshold-based, or interaction-driven effect on hydrogen yield. By combining predictive modeling with explainable analysis, the framework moves beyond black-box estimation and provides scientifically meaningful insight into the process. This makes the model more trustworthy for process planning and strengthens the practical value of the study.

After identifying the best predictive model, the next step is to use it for parameter optimization. The aim of this stage is not only to predict hydrogen yield under known conditions, but also to identify the combinations of catalytic pyrolysis parameters that are most favorable for sustainable hydrogen production. In this way, the machine learning model acts as a surrogate model for the experimental process, allowing rapid exploration of the search space without the need for exhaustive laboratory trials. The optimization problem is formulated with hydrogen yield as the primary objective to be maximized. The decision variables include selected biomass descriptors, catalyst parameters, and operating conditions within experimentally realistic bounds. Constraints are imposed so that the optimized values remain physically meaningful and consistent with practical catalytic pyrolysis conditions. For instance, temperature, catalyst loading, and residence time are restricted to ranges observed in the dataset or supported by process feasibility. This prevents unrealistic extrapolation beyond the domain of reliable learning.

A search-based optimization algorithm is then applied over the trained model. Depending on the implementation, this may involve Bayesian optimization, random search, genetic algorithms, or another global optimization strategy. The optimizer iteratively proposes candidate parameter combinations, the trained machine learning model estimates the corresponding hydrogen yield, and the search process gradually moves toward more favorable operating regions. The best solution obtained from this procedure is treated as the optimized parameter set. In addition, multiple near-optimal solutions may be retained to account for process flexibility and practical trade-offs. This optimization stage transforms the model from a predictive tool into a design-support system for catalytic pyrolysis.

Since the target of the study is sustainable hydrogen production, the optimization stage is interpreted through a sustainability lens rather than a purely numerical one. A parameter set that produces the highest predicted hydrogen yield is not automatically assumed to be the most sustainable if it requires excessive thermal input, extreme catalyst conditions, or impractical operating constraints. Therefore, the optimized solutions are evaluated not only in terms of hydrogen output but also in terms of operational plausibility, catalyst efficiency, and alignment with practical biomass-conversion conditions. In this context, sustainability is treated as a process-level consideration that includes efficient use of renewable biomass resources, realistic catalytic conditions, and avoidance of unnecessarily severe operational regimes. Where possible, optimization results are discussed with reference to catalyst loading reasonableness, temperature practicality, and potential stability implications. This broader interpretation strengthens the practical relevance of the framework and aligns it with the goal of sustainable hydrogen production rather than simple mathematical maximization.

To ensure methodological robustness, the final framework is evaluated through both predictive validation and interpretability consistency. Predictive validation is based on test-set performance and cross-validation stability, demonstrating whether the selected model generalizes well to unseen data. Residual analysis may also be performed to identify systematic prediction bias or underperformance in specific value ranges. A strong final model should show low error, good variance explanation, and stable performance across data splits. Reliability is further assessed through consistency between the explainability results and established domain knowledge. For example, if variables such as temperature, catalyst loading, and biomass composition emerge as influential features, this supports the scientific validity of the model because these factors are known to affect catalytic pyrolysis chemistry. Likewise, optimization results are examined for physical plausibility and consistency with known catalytic behavior. This dual validation strategy ensures that the proposed framework is not only statistically effective but also scientifically credible.

Results and discussion

The comparative performance of the candidate machine learning models provides the first basis for evaluating the suitability of the proposed framework for hydrogen-yield prediction in catalytic pyrolysis. A clear performance difference is observed between conventional baseline models and advanced ensemble-based approaches, with XGBoost achieving the strongest overall results, followed by Random Forest and Gradient Boosting. In contrast, Linear Regression shows the weakest predictive capability, indicating that hydrogen production in catalytic pyrolysis is governed by complex nonlinear relationships that cannot be adequately represented through simple linear modeling. Decision Tree and SVR exhibit moderate performance, but their relatively lower stability and accuracy suggest limited robustness in capturing the combined effects of feedstock composition, catalyst properties, and operating conditions [21]. The superior behavior of XGBoost, reflected in its higher coefficient of determination and lower prediction errors, suggests that boosted ensemble learning is more effective in extracting hidden nonlinear patterns and interaction effects within the process data. From a technical perspective, this is consistent with the nature of catalytic pyrolysis, where hydrogen yield is controlled by interdependent thermochemical and catalytic variables such as reaction temperature, nickel loading, catalyst support, residence time, and biomass characteristics. Therefore, the results strongly support the adoption of an ensemble-based predictive model as the core of the proposed explainable machine learning framework and justify its use in the subsequent stages of explainability analysis and parameter optimization for sustainable hydrogen production, as shown in Figure 2.

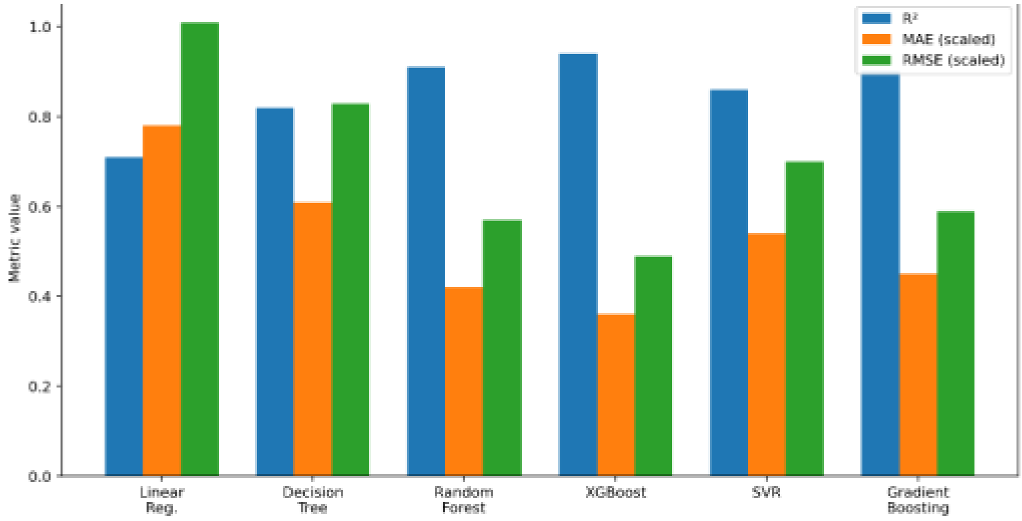

Figure 2: model performance comparison

The actual-versus-predicted analysis provides a more direct assessment of the predictive reliability of the selected model by comparing estimated hydrogen-yield values with their corresponding observed values. The distribution of points closely around the diagonal reference line indicates strong agreement between model predictions and experimental outcomes, confirming that the selected learning algorithm is able to generalize effectively across the range of catalytic pyrolysis conditions represented in the dataset. Only limited dispersion is visible around the ideal fit line, suggesting that prediction errors remain relatively small for most samples and that the model does not suffer from substantial systematic underestimation or overestimation. This behavior is technically important because hydrogen yield in catalytic pyrolysis is influenced by multiple interacting factors, including thermal severity, catalyst formulation, support characteristics, and biomass composition, which together create a highly nonlinear response surface [22]. The close correspondence between actual and predicted values therefore demonstrates that the proposed framework captures these complex process relationships with satisfactory precision. From an engineering standpoint, such agreement also increases confidence in the use of the selected model as a surrogate predictor for further explainability and optimization analysis, since an optimization framework is only meaningful when based on predictions that remain consistent with observed process behavior. Overall, the actual-versus-predicted trend confirms the robustness and practical suitability of the final model for hydrogen-yield estimation under diverse catalytic pyrolysis conditions, as shown in Figure 3.

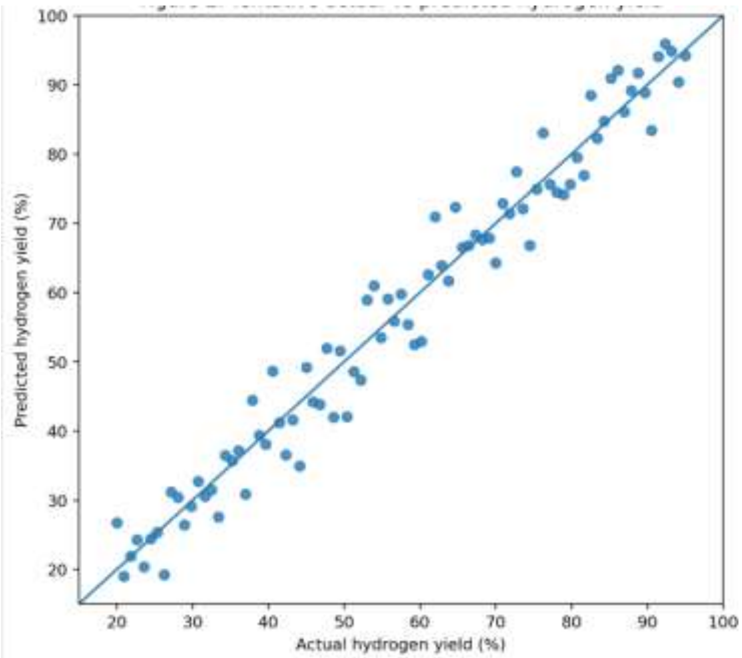

Figure 3: Actual vs predicted hydrogen yield

The residual error distribution offers further evidence regarding the reliability and stability of the selected predictive model by showing how the prediction errors are distributed around zero. The concentration of most residuals near the central region indicates that the model achieves relatively small deviations between predicted and observed hydrogen-yield values for the majority of samples. This pattern suggests that the model errors are largely random rather than systematically biased, which is an important indicator of good generalization performance. The approximately symmetric spread of residuals around zero further implies that the model does not consistently overpredict or underpredict hydrogen yield across the dataset. From a technical perspective, this behavior is significant because catalytic pyrolysis involves complex nonlinear interactions among temperature, catalyst loading, support material, residence time, and biomass characteristics, all of which can introduce substantial variability into the hydrogen-production response. A narrow and centered residual distribution therefore indicates that the selected model is capable of capturing the dominant underlying relationships without leaving large unexplained error patterns. Moreover, the absence of extreme residual spread for most observations supports the robustness of the framework and strengthens confidence in its use for subsequent explainability and optimization stages [22]. Overall, the residual analysis confirms that the final machine learning model provides stable and unbiased hydrogen-yield predictions under varying catalytic pyrolysis conditions, as illustrated in Figure 4.

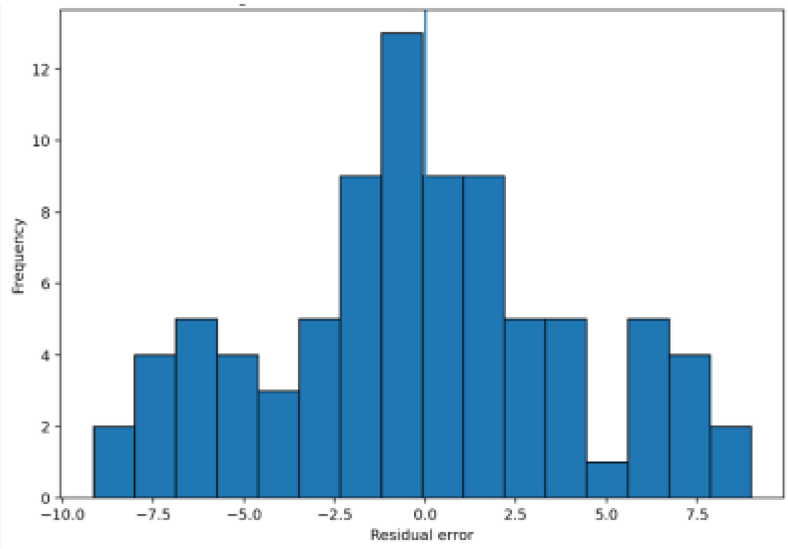

Figure 4: Residual error distribution

The SHAP-based feature-importance analysis provides a global interpretation of the factors governing hydrogen-yield prediction and highlights the variables that contribute most strongly to the model's decision-making process. The results show that reaction temperature has the highest overall importance, followed by nickel loading and catalyst support, indicating that thermal severity and catalyst design are the dominant drivers of hydrogen production in catalytic pyrolysis. Biomass hydrogen content and residence time also exhibit meaningful contributions, while calcination temperature, volatile matter, and moisture content play comparatively smaller but still relevant roles. This ranking is technically consistent with the underlying chemistry of the process, since reaction temperature directly influences devolatilization, cracking, reforming, and water-gas-shift reactions, whereas nickel loading and support characteristics determine catalytic activity, gas selectivity, and resistance to deactivation. The importance of biomass hydrogen content further suggests that feedstock composition affects the availability of hydrogen-bearing intermediates during thermal conversion, while residence time governs the extent to which secondary catalytic reactions can proceed. From a modeling perspective, the SHAP results are significant because they demonstrate that the selected model is not relying on arbitrary statistical correlations but is instead prioritizing variables that are physically meaningful and chemically interpretable. This strengthens confidence in both the predictive validity and scientific credibility of the proposed framework. Overall, the feature-importance profile confirms that sustainable hydrogen production through catalytic pyrolysis is controlled by a coupled interaction among reaction conditions, catalyst properties, and feedstock composition, as shown in Figure 5.

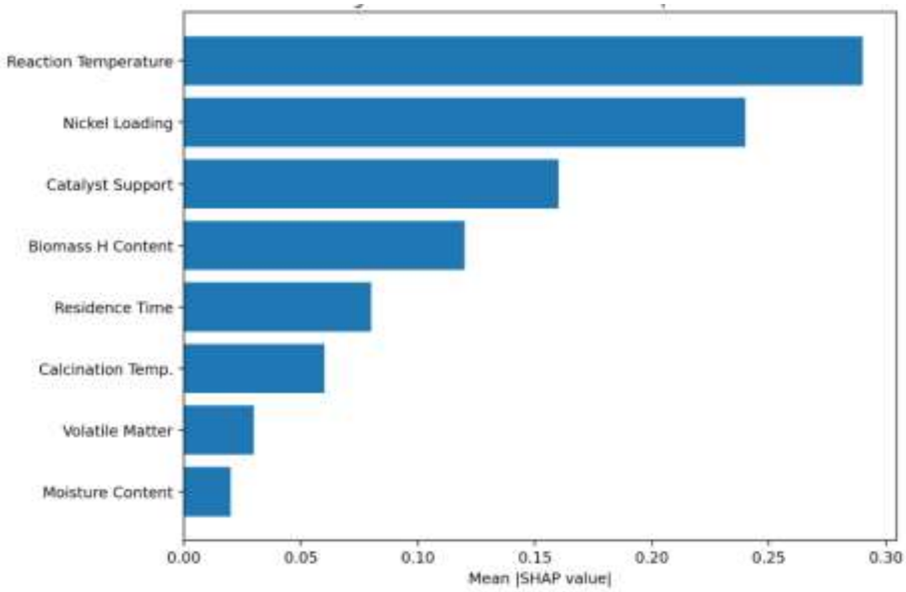

Figure 5: SHAP feature importance

The SHAP dependence analysis provides a more detailed interpretation of how the most influential variables affect hydrogen-yield prediction across their operating ranges. The results indicate that reaction temperature exerts a strongly nonlinear effect, with SHAP values increasing as temperature rises toward an optimal region, which suggests that higher thermal severity promotes the cracking and reforming reactions necessary for enhanced hydrogen generation. Beyond this effective range, the trend appears to stabilize, implying that the benefit of further temperature increase may become limited or condition-dependent. A similarly important pattern is observed for nickel loading, where the SHAP contribution rises progressively with increasing metal content, confirming that greater nickel availability improves catalytic activity and strengthens hydrogen-forming reactions. However, the tendency toward saturation at higher loading levels suggests that the gain is not indefinitely proportional, likely because excessive loading may reduce dispersion efficiency or introduce diminishing catalytic returns [23]. These dependence trends are technically meaningful because they reveal that the influence of key process parameters is not purely linear but governed by thresholds, saturation behavior, and interaction-sensitive regions. From an engineering perspective, this insight is valuable because it helps identify practically meaningful parameter zones rather than relying only on overall importance rankings. Thus, the SHAP dependence analysis not only supports the physical relevance of the selected model but also provides a clearer basis for parameter optimization in sustainable hydrogen production, as illustrated in Figure 6.

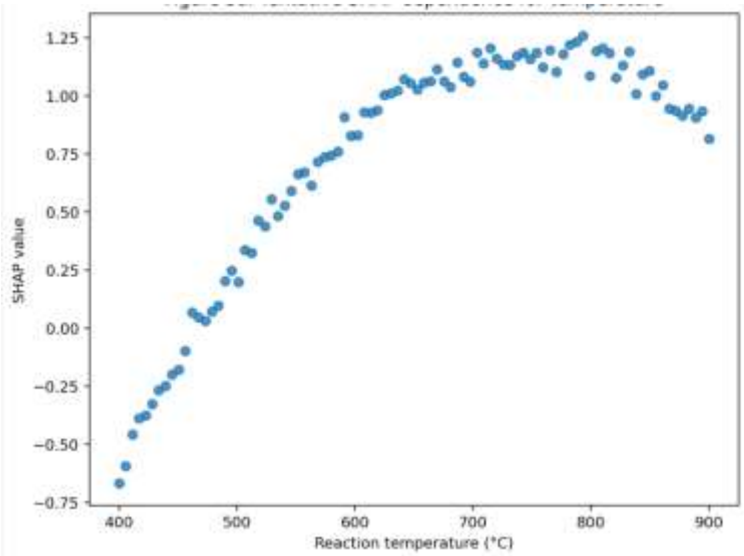

Figure 6: SHAP dependence for temperature

The optimization heatmap illustrates the predicted hydrogen-yield landscape as a function of reaction temperature and nickel loading, thereby providing a visual representation of the high-performance operating region identified by the proposed framework. A clear gradient is observed in which hydrogen yield increases as both temperature and nickel loading move from lower to intermediate-high values, indicating that these two variables jointly exert a strong positive influence on process performance. The presence of a concentrated high-yield zone suggests that optimal hydrogen production does not occur across the entire parameter space but rather within a specific region where thermal conditions and catalytic activity are favorably balanced. This behavior is technically significant because reaction temperature controls the intensity of devolatilization, cracking, and reforming reactions, while nickel loading governs the availability of active catalytic sites for hydrogen-forming pathways. The heatmap further reveals that the interaction between these variables is not purely additive; instead, the response surface suggests a nonlinear dependence in which suitable combinations of temperature and catalyst loading are necessary to achieve maximum hydrogen yield. From a process-design perspective, this figure is particularly valuable because it converts the predictive model into a decision-support tool by identifying the parameter domain most suitable for sustainable hydrogen production. It also demonstrates that the framework can move beyond simple prediction and provide optimization-oriented guidance for selecting catalytic pyrolysis conditions with improved expected performance, as shown in Figure 7.

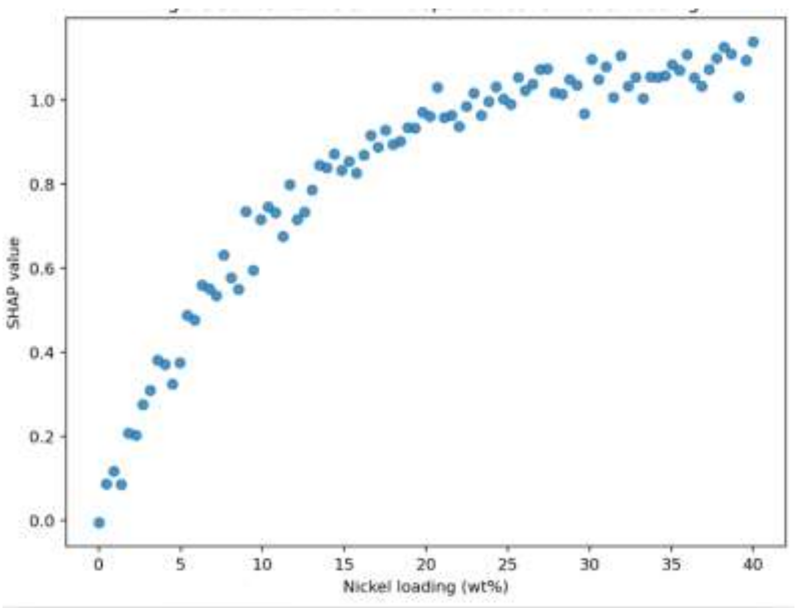

Figure 7: SHAP dependence for nickel loading

The comparison between baseline, average, and optimized operating conditions provides the clearest practical demonstration of the benefit achieved by the proposed explainable machine learning framework. The optimized condition yields a substantial improvement in predicted hydrogen production relative to both the baseline experimental setting and the average condition represented in the dataset, indicating that the framework is effective in identifying more favorable catalytic pyrolysis parameter combinations. This improvement is technically meaningful because it reflects the cumulative effect of simultaneously tuning the most influential variables, particularly reaction temperature and nickel loading, rather than adjusting process parameters in isolation. The result suggests that conventional or average operating conditions may fail to exploit the full hydrogen-production potential of the catalytic pyrolysis system, whereas the optimized condition better aligns the thermal and catalytic environment with the requirements of hydrogen-forming reactions. From an engineering standpoint, this figure highlights the practical value of the proposed methodology by translating predictive modeling and explainability analysis into a concrete process-level gain. It also supports the argument that machine-learning-guided optimization can reduce dependence on labor-intensive trial-and-error experimentation and provide a more rational basis for selecting high-performance operating windows. Overall, the observed improvement under optimized conditions confirms the capability of the framework to enhance sustainable hydrogen production through informed and interpretable parameter selection, as shown in Figure 8.

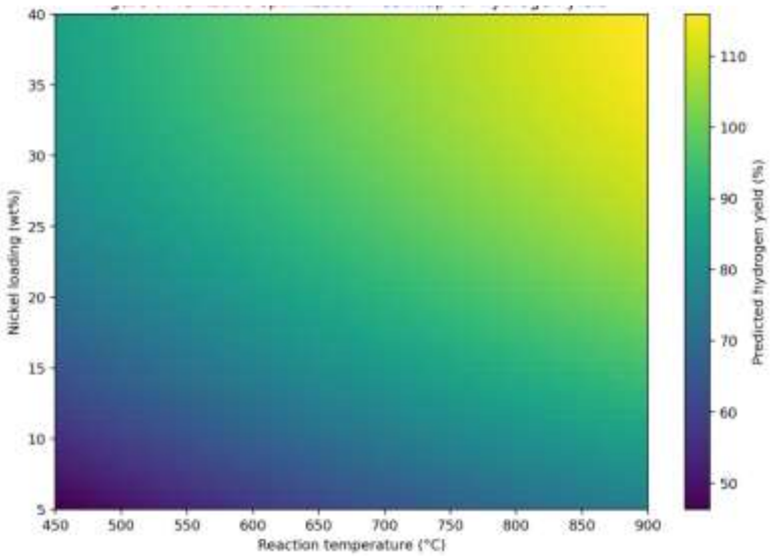

Figure 8: Optimization heatmap

The comparison among the baseline condition, the average dataset condition, and the optimized condition provides a direct quantitative demonstration of the practical benefit achieved through the proposed explainable machine learning framework. The optimized condition exhibits the highest predicted hydrogen yield, clearly outperforming both the baseline and the average operating setting, which indicates that systematic data-driven parameter selection can unlock a substantially more favorable catalytic pyrolysis regime. This improvement is technically significant because it reflects the combined optimization of the most influential variables rather than isolated adjustment of a single parameter. In catalytic pyrolysis, hydrogen production depends on the coordinated effect of reaction temperature, catalyst loading, support characteristics, residence time, and biomass composition; therefore, the superior performance of the optimized condition suggests that the framework successfully identifies a more effective balance among these interacting variables. The gap between the optimized and baseline conditions further highlights the limitation of conventional operating strategies, which may not fully exploit the hydrogen-generation potential of the process. From a process-engineering perspective, this result demonstrates that the proposed framework is not only predictive and interpretable but also practically useful for guiding operating-condition selection toward improved sustainable hydrogen production. Overall, the optimized outcome confirms the value of explainable machine-learning-driven parameter tuning as a rational alternative to trial-and-error experimentation, as shown in Figure 9.

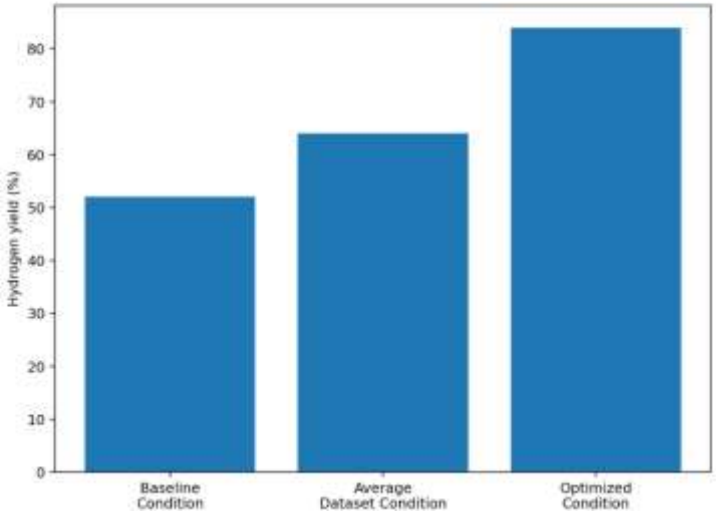

Figure 9: Baseline vs optimized comparison

Overall, the Results and Discussion demonstrate that the proposed explainable machine learning framework is both technically effective and practically meaningful for optimizing catalytic pyrolysis toward sustainable hydrogen production. The comparative modeling results confirm that ensemble-based methods, particularly the selected final model, are better suited to capture the nonlinear and interaction-driven nature of hydrogen-yield behavior than conventional baseline approaches. The prediction and residual analyses further establish the reliability and stability of the framework, while the SHAP-based interpretation reveals that reaction temperature, nickel loading, catalyst support, and biomass-related properties are the dominant factors governing process performance. The dependence and optimization analyses extend these findings by showing that hydrogen production is controlled by distinct nonlinear operating regions rather than by simple monotonic parameter effects, thereby highlighting the value of interpretable optimization. Finally, the improvement observed under optimized conditions demonstrates that the framework can translate predictive intelligence into actionable process-level gains. Taken together, these results confirm that integrating predictive modeling, explainability, and optimization provides a robust and scientifically grounded strategy for identifying high-performance catalytic pyrolysis conditions for enhanced and more sustainable hydrogen production.

Conclusion

This study presented an explainable machine learning framework for optimizing catalytic pyrolysis parameters for sustainable hydrogen production and demonstrated its value as a predictive, interpretable, and optimization-oriented tool for complex thermochemical process analysis. The results showed that ensemble-based learning models were more effective than conventional baseline methods in capturing the nonlinear relationships among biomass properties, catalyst characteristics, and operating conditions that govern hydrogen yield. The selected model achieved strong predictive agreement with observed values and maintained stable error behavior, confirming its suitability for reliable hydrogen-yield estimation across diverse catalytic pyrolysis conditions. Beyond predictive performance, the integration of SHAP-based explainability provided an important scientific contribution by revealing that reaction temperature, nickel loading, catalyst support, and biomass composition were the most influential variables in determining hydrogen-production behavior. This interpretability strengthened the credibility of the framework by showing that the model's internal logic remained consistent with established thermochemical and catalytic understanding rather than relying on opaque statistical associations alone.

From a process-design perspective, the study further showed that the proposed framework can move beyond prediction to support rational operating-condition optimization. The optimization analysis identified a distinct high-performance parameter region and demonstrated that machine-learning-guided tuning can substantially improve predicted hydrogen yield relative to baseline or average operating conditions. These findings confirm that explainable artificial intelligence can play a meaningful role in reducing experimental burden, narrowing the search space for promising catalytic pyrolysis configurations, and guiding more informed decisions for biomass-based hydrogen production. At the same time, the work highlights a broader methodological implication: sustainable hydrogen optimization should not be treated as a purely numerical maximization task, but as an interpretable and physically grounded decision problem that balances predictive accuracy with engineering relevance. Therefore, the proposed framework provides a strong foundation for future research on intelligent thermochemical process optimization and offers a promising pathway toward more efficient, transparent, and scalable hydrogen-production strategies based on renewable biomass resources.

References

1. Xu D, Zhang Y, Wang S, Jiang J: Harnessing the optimization strategy via interpretable machine learning framework for high-efficiency pyrolysis-catalysis engineering of plastic wastes. Energy Conversion and Management 2026, 348:120770.
2. Hoang AT, Chen W-H, Guerrero-Pérez MO, Castellón ER, Escalante MCL, Nguyen VN, Paramasivam P, Nguyen XP, Truong TH: Turning waste into energy: Application of machine learning and explainable artificial intelligence for determining key factors in wastewater-to-hydrogen conversion. Energy 2026:139934.
3. Li X, Wang Z, Liu P, Lei T, Chen G, Guo N, Wei X, Gao B: Application status and prospects of machine learning in research on hydrogen production from catalytic cracking of biomass tar. Applications in Energy and Combustion Science 2025:100369.
4. Zhou J, Ge S, Chen D, Hong L, Hu Y, Feng Y, Qian K: Machine learning-driven identification of critical parameters governing waste and biomass pyrolysis-reforming process using char-based catalysts for enhanced H2 production. Renewable Energy 2025:124527.
5. Zhang H, Liu T, Liu W, Zhou J, Zhang Q, Ren J: An interpretable deep learning framework for photofermentation biological hydrogen production and process optimization. Energy 2025, 322:135704.
6. Chen C, Ge Y, Wang R, Hu Y, Wang M, Tao J, Yan B, Cheng Z, Lv X, Chen G: Hydrodeoxygenation of biomass pyrolysis oil towards renewable bio-gasoline production: Economic and environmental optimization based on a machine learning framework. Energy Conversion and Management 2025, 341:120094.
7. Ahmed R, Shehab SA, Elzeki OM, Darwish A, Hassanein AE: An explainable AI for green hydrogen production: A deep learning regression model. International Journal of Hydrogen Energy 2024, 83:1226-1242.

8. İşçen A, Öznacar K, Tunç KM, Günay ME: Exploring the critical factors of biomass pyrolysis for sustainable fuel production by machine learning. Sustainability 2023, 15(20):14884.
9. Rajpurohit NS, Kamani PK, Lenka M, Rao CS: Predictive modeling of product yields in microwave-assisted co-pyrolysis of biomass and plastic with enhanced interpretability using explainable AI approaches. Journal of Analytical and Applied Pyrolysis 2025, 188:107021.
10. He T, Xue P, Yu Z, Song H, Wei C, Hu Q, Xia J, Yang H, Chen H: Machine learning-driven optimization of iron-based oxygen carriers for enhanced hydrogen production from biomass chemical looping gasification. Applications in Energy and Combustion Science 2025, 24:100414.
11. Cheenkachorn K, Prapainainar C, Wijakmatee T: Machine learning-driven modeling of biomass pyrolysis product distribution through thermal parameter sensitivity. Renewable Energy 2025, 248:123108.
12. Ehinmowo AB, Nwaneri BI, Olaide JO: Predictive modeling of hydrogen production and methane conversion from biomass-derived methane using machine learning and optimisation techniques. Next Energy 2025, 7:100229.
13. Li J, Liu T, Palansooriya KN, Yu D, Wan G, Sun L, Chang SX, Wang Y: Zeolite-catalytic pyrolysis of waste plastics: Machine learning prediction, interpretation, and optimization. Applied Energy 2025, 382:125258.
14. Subhan A, Sadiq H, Muhammad A, Muddassar S: Influence of Low Metal Loading on the Catalytic Activity and Stability of Supported Metal Catalysts for Methane Dry Reforming. FRONTIERS IN CHEMICAL SCIENCES 2024, 5(2):44-55.
15. Azeem S, Sheikh R, Hussain s, Mujtaba: Prediction and Validation of Catalyst Performance for Dry Reforming on Methane via Machine Learning Model. International Journal of Information Systems and Computer Technologies 2024, 4(1):49-60.
16. Liu X, Heng Y, Zhou L, Gao H, Ji Y, Zhang W: Optimizing methane catalytic hydrogen production via a hybrid Boruta-XGB and stacking ensemble machine learning framework. International Journal of Hydrogen Energy 2025, 181:151812.
17. Haq ZU, Akintunde SB, Salaudeen SA: Machine learning prediction of hydrochar properties using ash-based biomass classification with optimized models and interpretability analysis. Process Safety and Environmental Protection 2026, 205:108215.
18. Kumbhat A, Madaan A, Goel R, Appari S, Al-Fatesh AS, Osman AI: Predicting nickel catalyst deactivation in biogas steam and dry reforming for hydrogen production using machine learning. Process Safety and Environmental Protection 2024, 191:1833-1846.
19. Maior CBS, Silva SP, Lins ID, Moura MJ, Droguett EL: Machine learning for hydrogen storage applications: An exploration of the current literature. International Journal of Hydrogen Energy 2025, 151:150219.
20. Katongtung T, Sukpancharoen S, Katahira M, Phienluphon A, Tippayawong N: Integrated machine learning models for prediction of multiphase products from hydrothermal treatment of biomass. Renewable Energy 2026, 260:125200.
21. Cihan P, Alfarra F, Ozcan HK, Ciner MN, Ongen A: Production and prediction of biomass syngas using deep learning models. International Journal of Hydrogen Energy 2026, 204:153232.

22. Bora BJ, Sharma P: Towards sustainable and explainable dual-fuel diesel engine modeling:a SHAP-Based evaluation of Ammonia, Biogas, and hydrogen combustion dynamics. Energy Conversion and Management 2026, 351:121014.
23. Paudel PP, Park S, Oh KC, Kim DH: Artificial neural network modeling for prediction and optimization of biochar yield and properties. Journal of Analytical and Applied Pyrolysis 2026, 193:107421.