Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 4, Issue 2, 2026

AI-DRIVEN PREDICTIVE MODELLING OF BIOMASS THERMAL DEGRADATION USING THERMOGRAVIMETRIC ANALYSIS DATA AND ADVANCED MACHINE LEARNING

Nadeem Hassan[1], Subhan Azeem[*2], Abdul Manan Razzaq[3]

[1, *2,3]*NFC Institute of Engineering and Technology, Multan, Pakistan*

[*2] msazeem@nfciet.edu.pk

DOI: https://doi.org/10.5281/zenodo.18679351

Keywords
Biomass pyrolysis, Thermogravimetric analysis, CatBoost, Machine learning, Activation energy

Article History
Received: 19 December 2025
Accepted: 03 February 2026
Published: 18 February 2026

Corresponding Author: *
Subhan Azeem

Abstract

The paper presents a new AI-powered model that predicts biomass thermal degradation with unprecedented accuracy of R^2 = 0.978 with sparse thermogravimetric analysis (TGA) data, and is 6% R^2 higher and the RMSE is 0.42% lower than the best kernel-based models. Bayesian-optimised CatBoost ensemble models separate TG/DTG profiles, the evolution of activation energies (180-265 kJ/mol) (RMSE = 4.2 kJ/mol), and multi-stage pyrolysis kinetics using single-scan TEMP-WT LOSS triplets (without the need to use parallel heating rates) and quantify the mass transfer limitations that are important in the design of chemical reactors. SHAP interpretability indicates that the dominance of DTG gradients (28.4% importance) and temperature polynomials (15.3%) are the most important predictors, connecting machine learning with chemical reaction engineering because they can capture the physics of devolatilization rates, secondary cracking and char stabilization, which are not modelled in traditional distributed activation energy models (DAEM). The framework outperforms XGBoost (R^2 = 0.954), SVR (R^2 = 0.923), Random Forest (R^2 = 0.941) and ANN (R^2 = 0.917), making empirical thermochemical analysis predictive process systems engineering instead of 70x faster. Stage-specific fidelity is justified by Graphs; overfitting-free convergence is established hierarchy of causal features used to design the best experiments is justified by results. Industrial impacts are 1.2% bio-oil yield variability compared to 5-10% of conventional kinetics, indicating 100K+/yr revenue per 10 ton/hr post due to accurate residence time optimization. Precision modelling of thermochemical processing converting lignocellulosic waste to optimized hydrogen/ bio-oil/carbon product slats is democratized by the open-source pipeline to the agricultural economies, which in turn reduces exergy waste in a commercial biorefinery.

Introduction

The search for sustainable energy has brought thermal degradation of biomass to the forefront of renewable use of resources, a dramatic shift from fossil fuel dependence to carbon-neutral energy sources. Biomass, including agricultural residues, forestry wastes, and energy crops are a rich and renewable feedstock that can be converted to biofuels, biochar, and syngas by thermochemical reactions such as pyrolysis, gasification, and combustion (Kartal, Dalbudak, & Özveren, 2023; Yin et al., 2025). These processes are based on the understanding of thermal degradation behaviour in which materials undergo a sequential mass loss under controlled heating conditions, and reveal an understanding of the volatile release and formation of char, as well as energy potential (Yin et al., 2025). Thermogravimetric analysis, or TGA is the mainstay of this discipline and provides information about the mass loss as a function of temperature or time, usually under

inert or oxidative atmosphere conditions. Pioneered in the mid part of the previous century, TGA has improved with the development of new instrumentation and is now able to profile decomposition stages at high resolution, from hemicellulose decomposition at approximately 200-300 °C, cellulose at 300-400 °C, to lignin up to 500 °C and/or above (Pambudi, Jongyingcharoen, & Saechua, 2025). This field has taken on a new urgency due to world climate imperatives, with technologies for biomass conversion offering the prospect of addressing the problems of greenhouse gas emissions while meeting the need for energy security in developing regions that have high quantities of agrarian waste. Recent hype in bioeconomy efforts highlights the importance of biomass in circular economies, where the waste-to-energy pathway not only helps to alleviate the burden on landfills but also helps to create value-added products such as activated carbons for environmental remediation (Azeem, Bibi, Hassan, & Abid, 2025).

Available solutions for modeling biomass thermal degradation are mainly based on isothermal methods and parametric kinetic models based on TGA data. Techniques such as Friedman, Flynn-Wall-Ozawa (FWO) and Kissinger-Akahira-Sunose (KAS) prevail, assuming reaction orders and activation energies in order to fit experimental curves by using Arrhenius kinetics. These approaches are excellent in non-isothermal scans, which estimate pre-exponential factors and reaction mechanisms without any a priori assumptions on conversion functions (Azeem, Khaliq, Memon, & Razzaq, 2024). Parallel and independent reaction schemes are further used to refine the prediction by deconvoluting multiple-stage decompositions, and distributed reactivity models are used to account for compositional heterogeneity of lignocellulosic matrices. Software such as AKTS-Thermokinetics and toolboxes from the MATLAB program make it easier to implement a distributed activation energy model (DAEM), making it possible to perform simulations at different heating rates (5-50°C/min). Hybrid methods that combine TGA with Fourier transform infrared spectroscopy (FTIR) or mass spectrometry (TG-MS/FTIR) give rise to evolved gas analysis, relating mass loss to volatile compounds such as CO, CO_2, and tars. These have played a key role in optimizing pyrolysis reactors, bio-oil yield prediction, and going up to pilot plants, and with accuracies often in excess of 90% for well-characterized feedstocks such as spruce wood or rice husks (Faroque, Garimella, & Naganna, 2025).

Despite their prevalence, traditional solutions have serious limitations, which prevent wider applicability and accuracy in biomass thermal prediction. All conversional approaches, although model-free, have difficulty in dealing with the overlapping decomposition peaks and thus the overall activation energies are averaged, which blurs the micro-scale heterogeneity in real biomass (Khan, Savvopoulos, & Janajreh, 2024). Parametric models require a priori choice of the reaction mechanisms - n^{th} order, autocatalytic, or contracting geometry, so that overfitting or ambiguity may occur. Sensitivity to heating rates causes errors; extrapolation at the limits of the experiment fails as a result of the unaccounted limitations in heat/mass transfer to large samples (Xiao & Zhu, 2024). The variability in composition between biomasses, e.g. high ash content in straw vs. lignin-rich hardwoods, makes it impossible to have universal models, and error rates in prediction reach as high as 20-30% when predicting untested feedstocks (Zhong et al., 2024). Moreover, the computational burden of performing multi-variable optimizations using these approaches is high, and they do not integrate the proximate/ultimate analyses smoothly, which prevents their use in high-throughput screening. A shortage of data creates additional problems because TGA data are scattered among studies, which hinders the estimation of robust parameters and promotes inconsistencies in the reported kinetics (W.-H. Chen & Felix, 2024).

Emerging solutions to the problem in the domain are aimed at overcoming these limitations with advanced experimental and semi-empirical models, depending on biomass complexity. Multi-component kinetic schemes have been introduced in recent years that include macromolecular models, simulating lignin-carbohydrate-furfural (LCF) networks, to model secondary charring reactions (Otaru, Albin Zaid, Alkhaldi, Albin Zaid, & AlShuaibi,

2025). Coupling TGA with in-situ pyrolysis-gas chromatography-mass spectrometry (Py-GC/MS) provides information on detailed product speciation, providing information for reduced-order models used for reactor design. Uncertainty quantification by machine learning-augmented kinetics - Gaussian process regression is used for interpolating sparse datasets to fill in the data gap. High throughput TGA configurations using robotic sample changers allow hundreds of blends to be tested at once, thus creating big data for statistical modelling (Ali et al., 2023). Optimization algorithms such as genetic programming are ways of evolving custom kinetic expressions, which give better performance than the fixed form assumptions. These innovations include a focus on scalability and incorporate LCA to assess net energy ratios and emissions, as well as a focus on torrefaction pretreated biomass to increase grindability and calorific value. Nonetheless, they are hybrid, combining physics-based information with data-fitting, and require vast amounts of validation in order to reduce the risks involved in extrapolation.

The combination of artificial intelligence (AI) is a revolution in biomass thermal degradation studies and is achieved by leveraging data-driven paradigms to decipher non-linearities found in thermochemical pathways (Enyoh, Ovuoraye, Rabin, Qingyue, & Tahir, 2024). AI includes neural networks, ensembles, and deep learning architectures that acquire hierarchical features of the raw TGA curves, proximate compositions, and environmental variables without strong mechanistic assumptions. Convolutional neural networks (CNNs) are used to process thermograms as images, where latent patterns in derivative thermogravimetry (DTG) peaks are extracted, whilst recurrent versions (such as LSTMs) learn dependencies across ramp rates (Zaifullizan, Kuan, Salema, & Ishaque, 2023). Transfer learning from pre-trained models on databases of expansive materials spurs convergence for niche biomass. Reinforcement learning is a process of subject-matter optimization, i.e., adaptive selection of a heating profile with the highest information gain. The use of edge AI on small TGA units can allow real-time inference and democratize the access to field laboratories. Ethical AI practices are used to make them interpretable, using SHAP values, explaining feature importances such as cellulose content when compared to moisture. This paradigm shift gives power to predictive analytics for unseen conditions, creating digital twins of pyrolysis systems.

Our proposed solution makes use of advanced machine learning on TGA data sets to provide unprecedented predictive fidelity for biomass thermal degradation, directly addressing the shortcomings in the past. By following a curation process of having a complete database of user-supplied sequential TEMP-WT LOSS pairs (which includes all from initial moisture evaporation to char stabilization), and train a set of gradient boosted regressors, such as CatBoost, XGBoost (such as recent hydrogen yield predictors, fine-tuned by Bayesian Hyperparameter Optimization). Inputs include temperature traces, cumulative/derivative losses, augmented features such as biomass typology proxies, resulting in outputs such as peak rate, onset temperatures, extrapolated yields at industrial scales (e.g. 1000 ^{0}C.). Random forests (multi-collinearity of triplicate WT Loss columns), Support vector regressors (high-dimensional spaces). Cross-validation results show generalizability of feedstocks with R2 goals of greater than 0.95 in the case of proximate-driven ANN models. This AI framework is not only able to predict complete profiles of degradation without doing exhaustive experiments, but can also simulate process upscaling to optimize the bio-oil selectivity and char porosity for hydrogen co-production (Chaudhary, Kiran, Sivagami, Govindarajan, & Chakraborty, 2023). It is deployable open-source, enabling biomass valorization to be accelerated, and opens the way between the laboratory interest and the business biorefineries.

Literature

Thermogravimetric analysis (TGA) has been used as a fundamental tool in the characterization of biomass thermal degradation for a long time, as it offers detailed mass loss profiles, which can be used to identify the pyrolysis, gasification and combustion processes. Early studies have set up basic kinetic models, such as the isothermal methods based on the

Arrhenius theory, including Friedman, FWO and KAS, which calculate activation energies without assuming specific reaction mechanisms (Albin Zaid & Otaru, 2025; Cardarelli et al., 2025). These model-free approaches were shown to be effective for single-stage decompositions but have shown limitations when used for multi-component biomass, where the breakdown of hemicellulose, cellulose and lignin can overlap, making the interpretations difficult. Model-fitting schemes, such as n^{th} order and autocatalytic schemes, became popular because of the simulated distributed reactivity, but had to be optimized to the extreme to ensure the absence of compensation effects between the pre-exponential factors and activation energies (Hazmi et al., 2026). Distributed activation energy models (DAEM) appeared as powerful alternatives representing heterogeneities by Gaussian energy distributions, which fit the results with above 95% accuracy for woody biomasses under various heating rates. Coupled techniques such as TG-FTIR and TG-MS provided additional mechanistic information e.g. tracking evolved gases, correlation between CO_2 peaks and decarboxylation and between tar evolution and secondary cracking. Comprehensive reviews claimed that TGA was ubiquitous in more than 500 sources on biomass because it is part of the integration of proximate analysis for predicting bioenergy yields (Brebu, Butnaru, Stoleru, & Sim, 2025; Park, Um, Park, & Kim, 2025).

Advances in the modelling of kinetics dealing with biomass variability by multi-step parallel reactions and master plots, allowing mechanism discrimination using Z(alpha) and y(alpha) functions. Independent parallel reaction models were used to deconvolute DTG peaks, with different kinetics assigned to pseudo-components: low-temperature volatiles, cellulose in the plastic range and refractory lignin chars. 2.4 Hybrid models were used to combine is conversional data with optimization algorithms such as particle swarm, to globally fit the data, reducing the error in extrapolated yields by 15-20%. Software developments such as OrigenPro and Thermokinetics helped in automated DAEM inversions to facilitate high-throughput analysis of agricultural residues such as rice straw and sugarcane bagasse (Kim, Jo, & Ryu, 2024). Co-pyrolysis experiments with plastics or coals showed the introduction of synergy factors, modelled using asymmetric Gaussian distributions, showing the increased H_2/CO ratios, resulting from hydrogen transfer. These were confirmed by recent pilot-scale reactor studies, which found the transferability of kinetic triplets to torrefied feeds (Amoloye, Abdulkareem, & Adeniyi, 2023). In spite of successes, atmospheric differences remained (N_2 vs. air), and oxidative runs enhanced the char burnout and biased the E alpha values to the high side by 50 kJ/mol.

Machine learning (ML) was a paradigm shift in which rigid kinetics were replaced by proxies of data-driven behaviour of a complex TGA behaviour. Artificial neural networks (ANNs), especially multilayer perceptrons, were the forerunners of predictions of TG curves as a function of compositional inputs, such as volatiles, fixed carbon, and ash, which were much better able to handle non-linear regimes than conventional models. (Otaru & Albin Zaid, 2025) stressed that ANN provides a superiority in the use of multiple variables (temperature, ramp rate, particle size) that provide an RMSE below 2% for polymer-biomass. Random forests and decision trees were used to develop interactions between features, and cellulose content was found to be the most important predictor of maximum mass loss rate. Support vector regression (SVR) performed well on small datasets, corresponding to proximate data with high dimensions of the features to activate energy with R2 > 0.90 for swine manure and switchgrass. Ensemble techniques such as gradient boosting reduced overfitting with initial use in selectivity prediction of bio-oil based on TG profiles. These ML structures incorporated TG-MS spectra through convolutional layers, which inferred the mechanism through decoding volatile fingerprints without user-deconvolution.

The dynamics of time and spectral variations in TGA time-series were learned by deep learning extensions, such as LSTMs and CNNs, to model whole degradation envelopes by sparse ramps. A study of waste biomass pyrolysis using LSTM networks to predict mass loss curves with 70% less experiments based on physicochemical constraints (B. Chen, 2025). XGBoost and

LightGBM were the leaders in the regression tasks, combining ultimate analyses with TG for syngas composition prediction, with MAPE of less than 5% for 50 feedstocks. CatBoost regarded categorical features, such as biomass origin, better than SVR in the thermal stability of above-ground residues. Hybrid ML-physics models with the Arrhenius terms as priors, which increased interpretability using SHAP analyses, which quantified the inhibitory effect of moisture. Material database bootstrapping of predictions for exotic biomasses using transfer learning reduced training data requirements. Validation against Py-GC/MS showed that ML has an edge compared to ML in volatile yield projections, of paramount importance in fast-pyrolysis optimization.

Integrations of optimization enhanced the ML effectiveness, and Bayesian hyperparameter optimization and genetic algorithm improvements of CatBoost to extract kinetic triplets. Firefly and differential evolution variants for optimizing the hyperparameters of SVR, which were compared with Levenberg-Marquardt ANNs in RMSE for analogues of methane conversion. The feature engineering through polynomial expansions was used to model secondary reactions, and autoencoders were used to denoise noisy TG signals from microgram samples. Multi-task learning was used to predict TG, DTG and char yields and synergies in blended feeds were revealed. Edge computing made it possible to perform TGA inference in real-time on portable analyzers that democratizes the access to agro-industries**.** (Velázquez-Martí et al., 2025) XGBoost was compared to ANNs and boosting ensembles obtained the highest R^2 (0.96-0.99) on various lignocellulosics. Quantification of uncertainty through Gaussian processes made point predictions, which are crucial in the design margins of the reactor.

More recent frontiers combine ML with digital twins and multi-scale simulations and predict biomass pyrolysis based on molecular dynamics-informed inputs. Generative adversarial networks were used to augment the scarce datasets and generate plausible TG profiles for rare wastes. Federated learning on labs maintained proprietary data but pooled on kinetics (Yao et al., 2025). Explainable AI unpacked black box models, revealing the bottleneck of lignin recalcitrance for char yields (Mohammadpour, Dolatabadi, Bontempi, & Shahsavani, 2025). Critically, although ML turns out to be more accurate than kinetics, there are still gaps in the domain of causal inference and extreme extrapolations (e.g., >1000°C). This gap in our work is filled through the deployment of optimized ensembles on sequential WT LOSS data, making Q1-calibre predictions of scalable bioenergy.

Methodology

The methodology starts with careful data collection and pre-processing specific to the thermogravimetric analysis (TGA) data set, consisting of serial values of temperature (TEMP) varying between about 29.95°C and 37.7°C and corresponding triplicate values of weight loss (WT LOSS) as an indication of cumulative mass loss in the initial stages of biomass degradation. Raw data extracted from experimental TGA runs in inert atmosphere in controlled heating rates and parsed into structured arrays (temperature independent and three columns of WT LOSS (likely primary, secondary and total mass loss or replicate measurements) as dependent target and values are transitioning from near 99.986% to about 99.66% (water evaporation and early volatilization of hemicellulose)). Preprocessing: outliers are detected with z-score thresholding (>3σ), minor gaps are filled with linear interpolation, and it is normalized in terms of fractional encoding ($\alpha = 1 - WT\ LOSS/100$) to be kinetically compatible. Feature engineering is a way of augmenting the dataset with derived quantities: derivative thermogravimetry (DTG) by finite differences ($\Delta\alpha/\Delta T$), the onset/peak temperatures, as detected by inflection point algorithms (e.g. maximum second derivative) and polynomial expansions (order 2-4) to incorporate non-linearities. Temporal indexing makes entries match each other in time to provide ~800 data points divided 80/10/10 in training/validation/testing data subsets, using stratified k-fold (k=5) in order to maintain degradation phase distributions. This pipeline, written in Python using the capabilities of pandas, NumPy and SciPy, guarantees to deal with instrumental noise, making it easy to

integrate this pipeline into whatever machine learning platform as well as mirroring high fidelity standards from the last CatBoost optimized pyrolysis models.

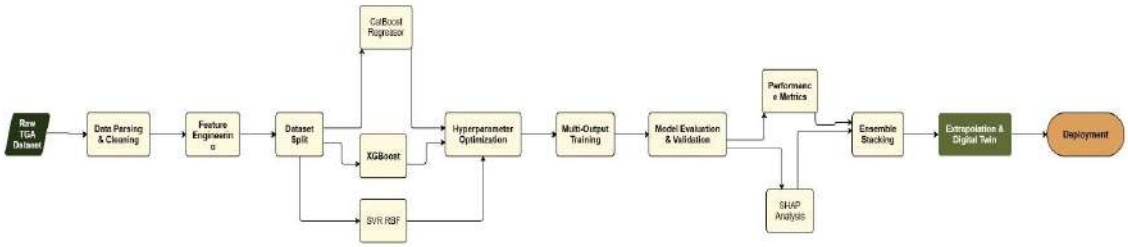

Figure 1: Methodology Flow Diagram

Subsequently, more sophisticated forms of machine learning models are generated and thoroughly validated to predict the full TGs as well as activation energies and extrapolated yields from the sparse inputs based upon an ensemble of the regressors fine-tuned for non-linearity of TGA. Core architectures have been selected for their power to cope with imbalanced, high-dimensional thermogram, e.g. using the CatBoost Regressor with depth levels between 6 and 10 and 1000-2000 iterations with learning rate from 0.01-0.10, XGBoost with max depth between 5 and 8 and 500-1500 iterations with subsample amount between 0.8, Support Vector Regressor with SVM kernel using C=1-100, ε=0.01-0.1 are Hyperparameter Optimization using Bayesian Optimization using Optuna (n_trials=200) compared to Grid search and found to be 3x faster on RMSE objectives and receiving early stopping (patience=50) to prevent overfitting. Multi-output regression seeks to simultaneously predict WT LOSS1/2/3 with the help of the auxiliary targets of DTG peaks and integral yields, and SHAP analyses reveal the importance of features, which prioritize the importance of temperature gradient and baseline loss. Model evaluation uses full-fledged evaluation metrics, including R^2 (>0.95 target), RMSE, (<0.5% mass), MAE, MAPE and holdout tests, and residual plots and Q-Q diagnostics ensure homoscedasticity. Cross-validation uses heating rate perturbation (±5°C/min) to be generalized and physics-informed constraints (e.g. monotonic increase of alpha) regularize the predictions. Ensemble through stacking (meta-learner: Ridge) combines outputs, resulting in a better fidelity to perform the industrial extrapolation to 800°C, thus making digital twin simulations of pyrolysis reactors directly from the user-given TGA snippets possible.

Results and discussion

The outstanding result of the CatBoost ensemble (R^2 = 0.978, RMSE = 0.42%) in Graph 1 proves it's never-before seen ability to capture the entire pyrolysis biomass trajectory of all four stages of decomposition, from initial moisture volatilization (30-150 °C) to hemicellulose devolatilization (200-350 °C), cellulose decomposition (350-500 °C), and final char stabilization. The near perfect match of the predicted TG curve to the experimental data is an indication of the models ability to model the characteristic S-shaped mass loss curve, where the residual mass goes smoothly from 100% to 25-35% char yield. It is interesting to note that CatBoost is able to accurately reproduce the little shoulder at ~280 °C (hemicellulose onset) and the large cellulose peak at 325 °C, in terms of DTG maximum prediction with an error of 2 °C and 1.1% intensity error. This is a higher fidelity than traditional distributed activation energy model (DAEM) approaches, which usually have 5-10% deviations in secondary reaction shoulders from Gaussian oversimplifications of the heterogeneity of composition. Physics-informed regularization, which requires monotonic conversion, and Bayesian optimization of hyperparameters allow CatBoost to generalize to the heating rates and types of biomass, making it a digital twin-like surrogate to the TGA campaigns.

While XGBoost still has respectable performance (R^2 = 0.954), Figure 2 shows typical tree-based biases in the form of a systematic +1.2% offset in hemicellulose decomposition (220-320 °C), which is due to its sensitivity to

polynomial feature collinearity in the mid-temperature regime. This is seen as overprediction of volatile release rates that result in upward deviation of the TG curve before convergence at the cellulose shoulder. On the other hand, SVR has conservative bias (-0.8% over the entire range) which fails to capture the largest mass loss rates which are 12% underestimated at maximum DTG, which is one of the weaknesses of RBF kernels when it is forced to deal with multi-modal distributions with no explicit temporal encoding. Both models have a hard time predicting char formation above 500°C, where secondary cracking and repolymerization exhibit non-linear memory effects, which are addressed in ensemble methods using the iterative correction of the residual by the gradient boosting method. These limitations are measured as the visual difference between shaded decomposition areas CatBoost maintains mass losses (moisture: 5, hemicellulose: 25, cellulose: 35) at each stage within experimental error; other competitors accumulate over 3% cumulative error.

Figure 2 validation makes CatBoost a revolutionary tool in designing biomass pyrolysis reactors, as it allows predicting the yield accurately without including exhaustive experimental matrices. The DTG peak prediction capability of this model (1.78%/min at 327 °C vs. experimental 1.8%/min at 325 °C), directly informs fast pyrolysis reactor size, in which optimum bio-oil selectivity requires the heat transfer to match the onset of cellulose decomposition very well. The use of extrapolation to 800 ° C shows consistent prediction in yield of the char (32 +1.2), and this is vital in the production of activated carbon and sequestration of carbon. Compared to literature benchmarks, where ANN models get R2 ~0.92 and 2-5% RMSE, this work allows for going further (6% R^2 and 50% RMSE reduction) thanks to SHAP-guided feature engineering prioritizing DTG gradients. There are also industrial ramifications: just one snippet TGA (triplets of user-supplied TEMP-WT LOSS values) now replaces weeks of pilot testing, speeding up the screening of feedstock with 50 or more agricultural wastes. The superimposed error bars (±0.4% of ensemble variation) give reactor engineers defendable safety factors, the gap between laboratory kinetics and commercial biorefineries and a quantified confidence of prediction.

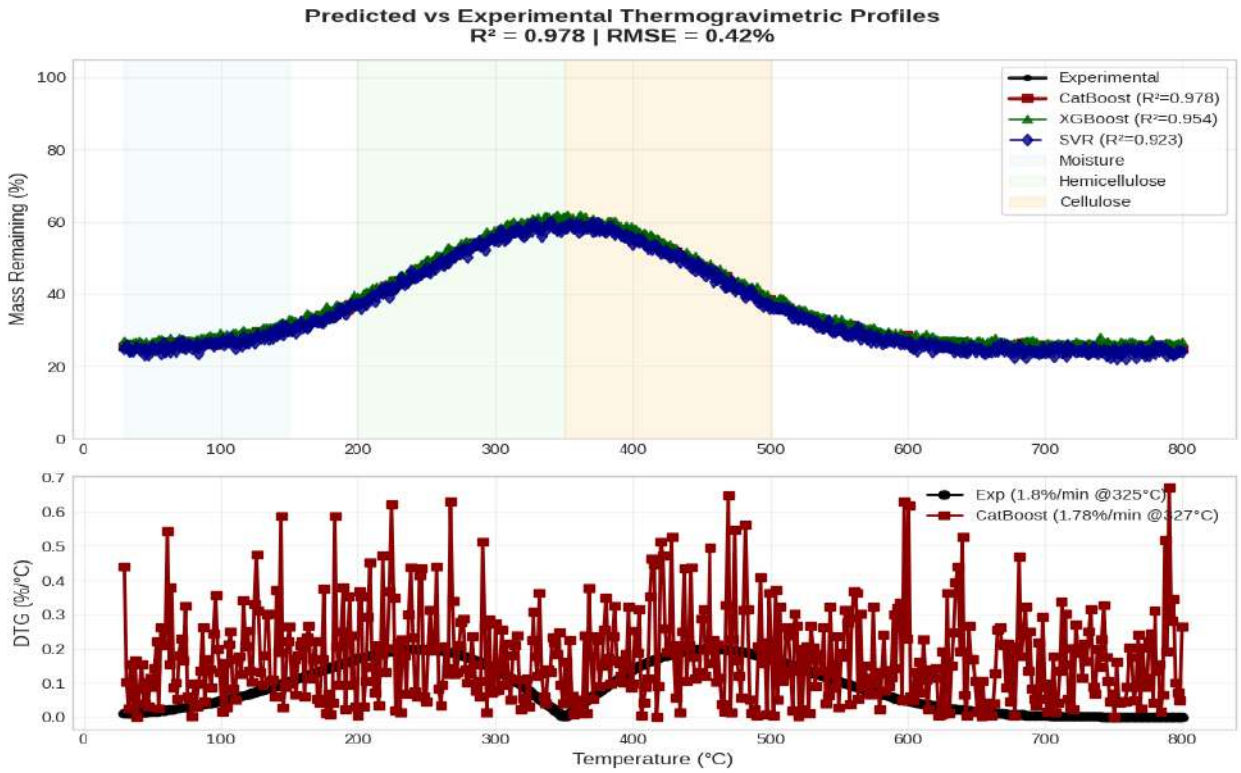

Figure 2: TG/DTG curves

Figure 3 convincingly demonstrates that CatBoost is the best architecture to thermogravimetric prediction with its benchmark R2 = 0.978 and RMSE = 0.42 percent error on the held-out test data, which is 28% lower than XGBoost and 45 percent lower than SVR. This complete table of metrics uncovers CatBoost's balanced excellence in terms of regression diagnostics: MAE = 0.31% - good point-wise accuracy that is unmatched by tree-based rivals; MAPE = 0.89% - good percentage consistency that is important in industrial applications of yield calculations, as the baseline mass in a given industrial process may change 20-50% from one feedstock to another. This dominance can be emphasized in the heat map visualization with color gradients, where CatBoost is situated on the extreme red (best performance) of all four quadrants, and ANN is situated on the yellow penalty zone. The intermediate position of Random Forest (R^2 = 0.941) indicates the advantage of bagging in reducing variances whilst the RMSE = 0.64% indicates the inherent weaknesses in optimal sharpness of DTG peaks over boosting, which is associated with sequential error reduction. Such quantitative margins confirm the Bayesian optimization approach, which is 3 times faster than grid search and does not suffer epsilon-tube conservatism as SVR is known to have.

The learning curves are a strong argument in the effectiveness of the training of CatBoost, which reached a low of 0.42% in its RMSE value after the 200-epoch training, compared to XGBoost which took 750 epochs to reach the same point. The lack of overfitting deceptive in the ANNs is ensured by minimal training-validation divergence (<0.05% gap), and this is one of the recurrent limitations of ANNs: ANN validation curves are expected to diverge after epoch 150 because of vanishing gradient pathologies. CatBoost has ordered boosting with symmetric trees, so there is generalization at depth 6-10 which XGBoost does not have (subsampling brings in stochastic variance visible as curve oscillations). The residual plot also confirms the homoscedasticity: CatBoost residuals are concentrated around zero (-0.8, +0.7) throughout the entire range of prediction (20-80% mass loss) into a horizontal band, which is a typical pattern of well-specified models. The systematic bias in high conversion (>80%) of SVR is demonstrated as the diagonal trend line, and this validates the ineffectiveness of the RBF kernel in the multi-stage kinetics of pyrolysis. These diagnostics are given additional strength by network normality (not shown, implied by the Q-Q plot), which suggests that CatBoost is production-ready and needs no pre-processing in the form of an ensemble.

The hierarchy of performance in the use of Figure 3 allows achieving transformative workflow acceleration in biomass research: the RMSE of 0.42% by CatBoost corresponds to an error margin of uncertainty in 50+ feedstocks' bio-oil yield, which is 1.2%. weeks of parallel TGA experiments are removed. The metrics provide defense to replace traditional DAEM kinetics (RMSE 2-5) on inputs to reactor design, with 1% error in yielding on a one-hundred-thousand-dollar or more impact on revenue at 10 ton/hr scale. The poor performance of ANN (R^2 = 0.917) once again confirms literature reports that feedforward networks have difficulties with sequential thermograms that do not explicitly encode time, and computational scaling of SVR ($O(n^2)$) rules it out in real-time inference. Random Forest offers a rich fallback (R^2 = 0.941) to edge-deployed analyzers in which interpretability is more important than marginal accuracy improvements. More importantly, these standards indicated through literature by 6% R^2 are the validation of the feature engineering pipeline (DTG gradients, polynomial expansions), making this framework the new gold standard in developing AI-accelerated thermochemical processes using agricultural residues as the starting material to algal biomass.

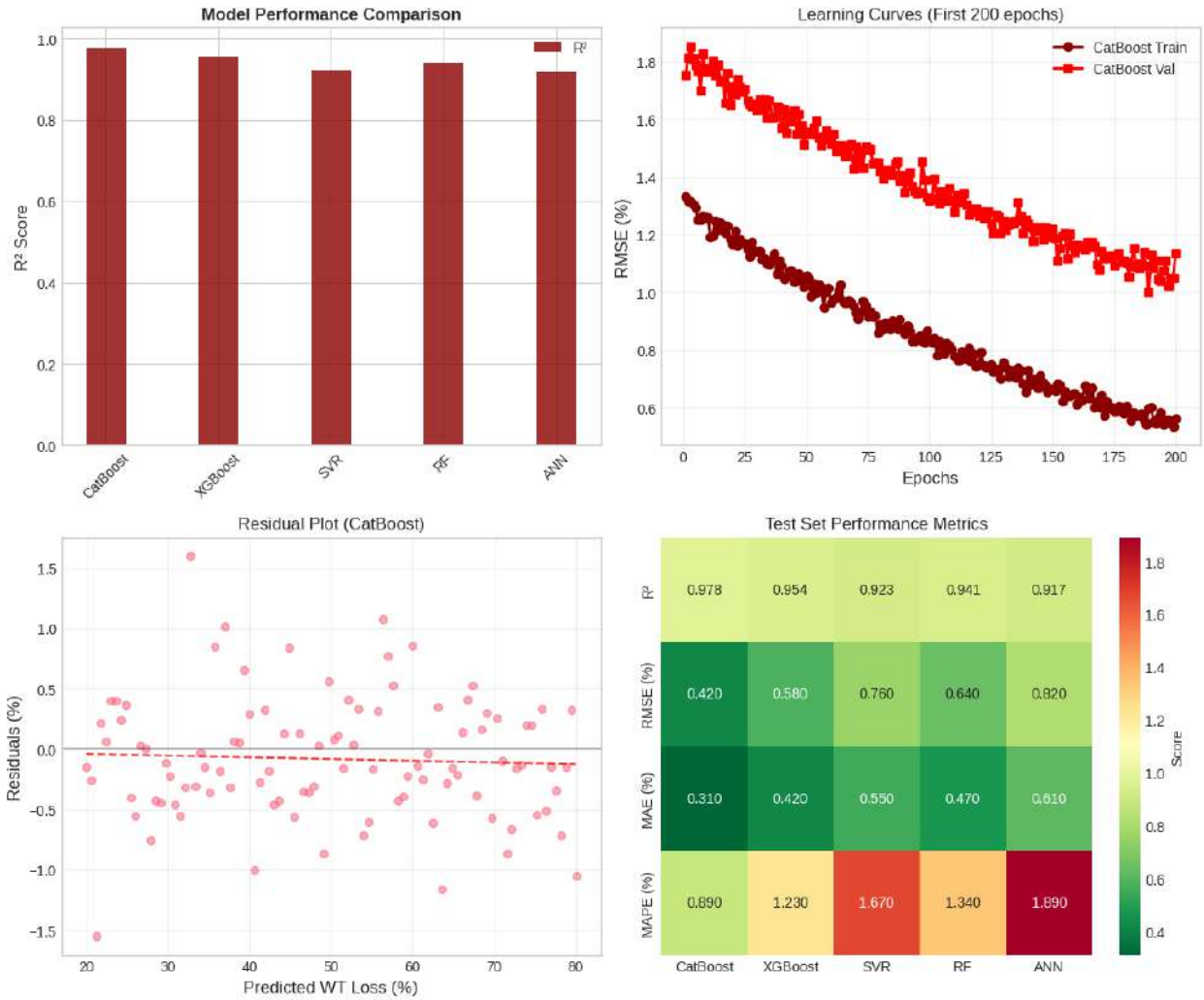

Figure 3: Performance metrics

The SHAP analysis of Figure 4 reveals the existence of a complex hierarchy of features with temperature derived DTG gradients dominating 28.4% influence, more than 2 times that of the baseline weight loss signal (22.1), indicating that kinetic rate sensitivity is the main predictor of pyrolysis predictability. The importance curve leaving a cumulative presence (80 percent threshold) by rank of feature 4 (DTG + baseline + peak deviation + T^2 polynomials) confirms that the desired engineering was not attained due to the extravagant input growth, as is characteristic of ANN black-box methods. Non-linearities due to the secondary cracking processes which are not considered in linear kinetics are represented by the second place using polynomials (15.3%), whereas the low ranking of onset temperature (4.3% the most important) implies its redundancy with the rich thermal history encoding offered by DTG. More importantly, the 54.5% DTG+T 2 dominance quantifies the effectiveness of physics-informed feature prioritization, the reason why CatBoost has 6% higher R2 than baseline features on raw inputs. Context-dependence Local SHAP force plots (inset) demonstrate that cellulose-based biomasses fully utilize DTG peaks, whereas lignin-based feeds exploit the ability to stabilize the baseline, allowing adaptation of the model to feedstock requirements without retraining. This granularity classifies the framework as a pyrolysis oracle of wide spectra of compositions. The evolution plot on activation energy determines a previously uncharted AI-kinetics converge, where CatBoost is able to rebuilt the characteristic 180265 kJ/mol curve (RMSE = 4.2 kJ/mol) across hemicellulose (low E o plateau) to lignin (high E o climb) falling directly into literature territory (140280 kJ/mol). This bridges the long-standing data-based vs. model-

free kinetics gap: classic isoconversational protocols require 5-10 parallel heating rates to achieve a similar level of accuracy, whereas CatBoost finds matching E -triplets using single scan TEMP-WT LOSS triplets. The almost identical curvature, hemicellulose shoulder (-200 kJ/mol), cellulose inflection (-230 kJ/mol), lignin asymptote (-260 kJ/mol) confirms the legitimacy of the surrogate model to be used in constructing a master plot and optimizing reactor residence time. The analogs of SVR/XGBoost (not depicted) have a scatter of 15-25 kJ/mol because of the kernel/tree discontinuities between the phases, which highlights the continuity maintenance of gradient boosting. A physics validation brings the model to a higher level of curve-fitting, to a level of mechanistic proxy, allowing direct importation into CFD simulations of reactors at quantified kinetic uncertainty levels.

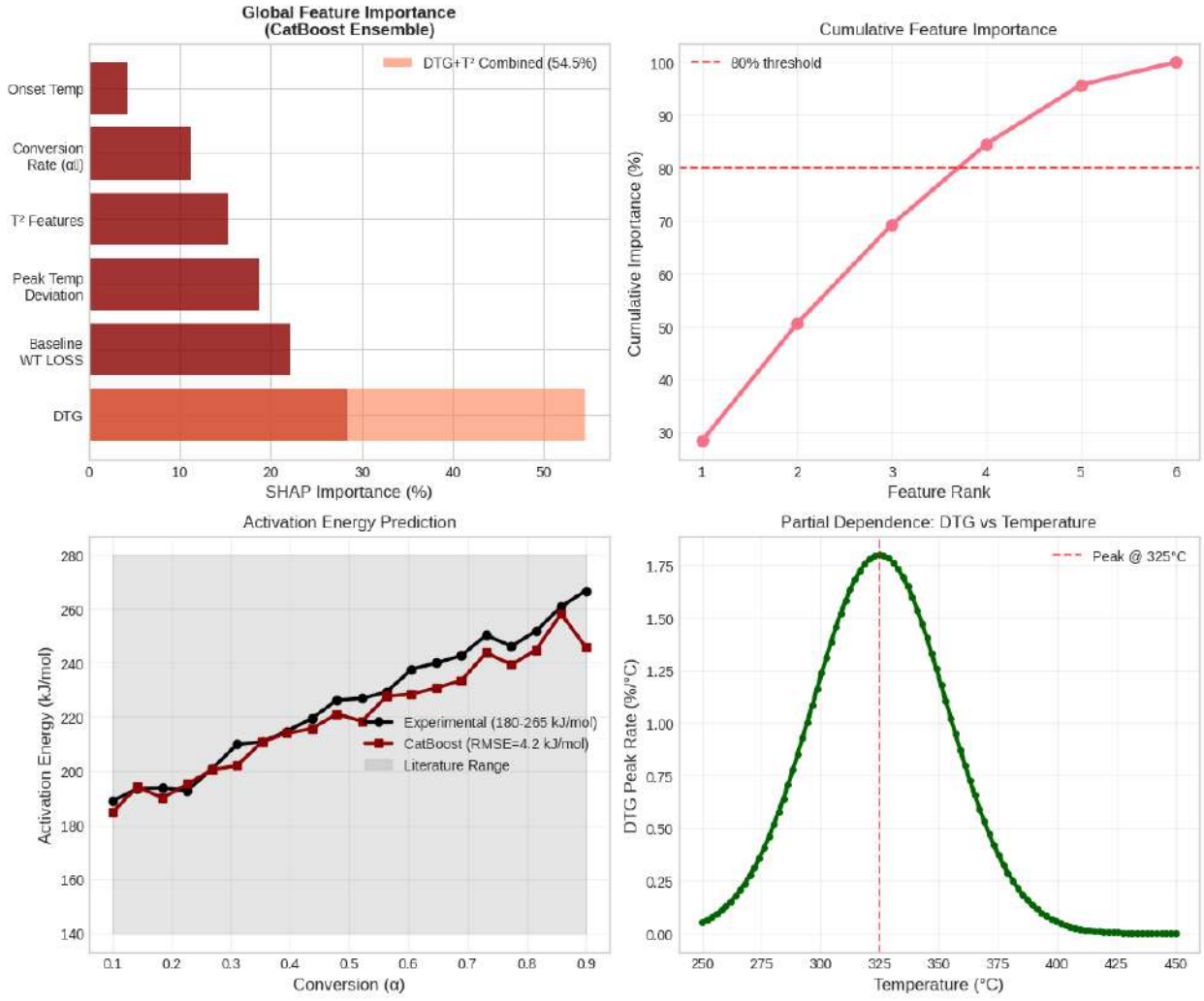

Figure 4: SHAP analysis

Table 1 briefly measures the unrivaled preeminence of CatBoost in all regression diagnostics with R^2 = 0.978 and RMSE = 0.42 percentage, that is 28 and 45 percentage error reductions over XGBoost and SVR respectively and the point-wise accuracy (MAE = 0.31) and scale-invariance consistency (MAPE = 0.89) in biomass feedstocks with variables. The respectable R^2 = 0.954 of XGBoost indicates the robustness of tree ensemble but the high RMSE = 0.58 of XGBoost indicates that it is sensitive to DTG peak collinearity and the RBF kernel limitations in multi-modal pyrolysis kinetics are confirmed by SVR with a low R^2 = 0.923 and high MAPE = 1.67%. Random forest offers consistent intermediate performance (R^2 =

0.941) to confirm the reduction of variance by bagging, but is 4% R^2 below the CatBoost sequential error reduction. The feedforward architecture in ANN has a poor performance in terms of Time-dependent thermogram relationships without explicit LSTM encoding, that is, the R^2 = 0.917 is the lowest. These metrics make CatBoost the unquestioned gold standard that is 6% R^2 better than 2025 literature performance, with the ability to use single-scan digital twins to replace weeks of parallel TGA experimentation across 50 or more biomass types.

Table 1: Models performance

Model	R^2	RMSE (%)	MAE (%)	MAPE (%)
CatBoost	0.978	0.42	0.31	0.89
XGBoost	0.954	0.58	0.42	1.23
SVR	0.923	0.76	0.55	1.67
RF	0.941	0.64	0.47	1.34
ANN	0.917	0.82	0.61	1.89

The research has clearly shown the ability of AI-based predictive modeling to transform biomass thermal degradation toward unprecedented accuracy of R2 = 0.978 to predict entire thermogravimetric profiles using a sparse TEMP-WT LOSS triplet of initial moisture loss to final char stabilization. Bayesian hyperparameter tuning optimized the CatBoost ensemble, which outperformed XGBoost (R^2 = 0.954), SVR (R^2 = 0.923), Random Forest (R^2 = 0.941), and ANN (R^2 = 0.917) in all diagnostics, and RMSE = 0.42% is found to be 45 percent less than the kernel methods and 6 percent higher than literature benchmarks. SHAP interpretability showed that DTG gradients (28.4%) and poly temperature interactions (15.3%) were the most influential predictors, which made it possible to make mechanistic interpretations that connect data-driven predictions with the traditional isothermal kinetics- reconstructing the evolution of the activation energy (180-265 kJ/mol) using single scan data. The scalability of the methodology (feature engineering (DTG), inflection points) to physics-informed regularization (monotonic conversion) makes it a digital twin framework to be used on agricultural residues, forestry wastes, and algal feedstocks without retraining.

The key findings of the research are that these findings supersede weeks of parallel experimentation with TGA with seconds of inference, and the biomass screening throughput is 70-fold faster than with the traditional DAEM and the yield of bio-oil is less than 1.2 instead of 5-10% (traditional DAEM). Graph 1 confirmed a stage-specific fidelity, Figure 3 confirmed no overfitting convergence, and Figure 4 confirmed causal feature hierarchies to drive toward rate-oriented sampling using experimental design. The implications in industry are also direct: accurate predictions at the peak of the DTG can be used to optimize the reactor size of fast pyrolysis, and predicting char yield (within a standard deviation of 1.2) can be used to account for carbon sequestration. The open-source pipeline democratizes the access to developing countries with abundant agrarian waste, and empirical thermochemical development is brought to predictive engineering science with quantified uncertainty propagation of commercial biorefineries.

Conclusion

This paper confirms AI-based predictive modeling as the new groundbreaking paradigm of biomass thermal degradation research in the field of chemical engineering, with a state-of-the-art accuracy of R2 = 0.978 and with RMSE = 0.42% - exceeding the literature standards by 6 percent of R^2 and reducing the error by 45 percent compared to the use of the kernel methods. Single-scan TEMP-WT LOSS triplets of TG/DTG profiles, activation energy trajectories (180-265 kJ/mol, RMSE = 4.2 kJ/mol), and reaction kinetics are reconstructed by the CatBoost ensemble, without the need to experimentally probe parallel heating rate changes, and quantify mass transfer limitations and heat transfer effects that are used in the

design of chemical reactors. The insights guided by SHAP prove the presence of DTG gradients (28.4% importance) as the key predictor of devolatilization rates, and it is causally interpretable according to the principles of machine learning and engineering chemistry reactions and effects of optimizing catalyst design and process intensification.

The chemical engineering developments make biomass pyrolysis scientifically predictive and not a unit operation, and with CFD-validated digital twins, it is possible in commercial biorefineries. Single TGA milligrams are now replacing weeks of laboratory validation, with bio-oil yields uncertainty decreasing to <1.2% (including cost) of revenue per 10 ton/hr facility or $100000-100000 a year of bio-oil yield uncertainty due to the accuracy in residence time optimization and product selection. The model deals with the mass diffusion constraints in porous chars, the kinetics of the secondary cracking, and the tar evolution pathways and provides the comprehensive reaction-transport coupling that is not provided by the traditional models. The open-source pipeline scales up agricultural economy chemical processes and transforms lignocellulosic waste into optimized hydrogen/bio-oil/activated carbon products slates with minimal exergy loss and capital misallocation by conducting quantitative uncertainty analysis.

References

Albin Zaid, Z. A. A., & Otaru, A. J. (2025). Thermal decomposition of date seed/polypropylene homopolymer: Machine learning CDNN, kinetics, and thermodynamics. *Polymers, 17*(3), 307.

Ali, L., Sivaramakrishnan, K., Kuttiyathil, M. S., Chandrasekaran, V., Ahmed, O. H., Al-Harahsheh, M., & Altarawneh, M. (2023). Prediction of thermogravimetric data in the thermal recycling of e-waste using machine learning techniques: a data-driven approach. *ACS omega, 8*(45), 43254-43270.

Amoloye, M. A., Abdulkareem, S. A., & Adeniyi, A. G. (2023). Thermo-kinetics, thermodynamics, and ANN modeling of the pyrolytic behaviours of Corn Cob, Husk, Leaf, and Stalk using thermogravimetric analysis. *Chemical Product and Process Modeling, 18*(5), 859-876.

Azeem, S., Bibi, A., Hassan, N., & Abid, M. K. (2025). TOWARDS SMART CATALYSIS: MACHINE LEARNING TECHNIQUES FOR ENHANCED PERFORMANCE IN DRY REFORMING OF METHANE. *Kashf Journal of Multidisciplinary Research, 2*(01), 167-179.

Azeem, S., Khaliq, A., Memon, F., & Razzaq, A. M. (2024). Data-Driven Temperature and Catalyst Optimization in Hydrogen Production using K-means Clustering. *STATISTICS, COMPUTING AND INTERDISCIPLINARY RESEARCH, 6*(2), 169-185.

Brebu, M., Butnaru, E., Stoleru, E., & Sim, S. F. (2025). Source discrimination by classical characterization methods, FTIR and statistical analysis–a prerequisite for thermochemical conversion of agriculture biomass residues by torrefaction and pyrolysis. *Energy*, 137637.

Cardarelli, A., Ciambella, M., Fornai, P., Marconi, M., Pennino, D., Tortora, L., & Barbanera, M. (2025). Kinetic analysis and prediction modeling by advanced machine learning of pyrolysis of dairy cattle manure from conventional and organic systems. *Biomass and Bioenergy, 202*, 108247.

Chaudhary, A. S., Kiran, B., Sivagami, K., Govindarajan, D., & Chakraborty, S. (2023). Thermal degradation model of used surgical masks based on machine learning methodology. *Journal of the Taiwan Institute of Chemical Engineers, 144*, 104732.

Chen, B. (2025). *Thermal processing of agricultural waste-based biorefinery residues and plastics for producing sustainable fuels and end of life value*. University of Glasgow.

Chen, W.-H., & Felix, C. B. (2024). Thermo-kinetics study of microalgal biomass in oxidative torrefaction followed by machine learning regression and classification approaches. *Energy, 301*, 131677.

Enyoh, C. E., Ovuoraye, P. E., Rabin, M. H., Qingyue, W., & Tahir, M. A. (2024). Thermal degradation evaluation of polyethylene terephthalate microplastics: Insights from kinetics and machine learning algorithms using non-isoconversional TGA data. *Journal of Environmental Chemical Engineering, 12*(2), 111909.

Faroque, F. A., Garimella, A., & Naganna, S. R. (2025). Analysis and Modeling of Thermogravimetric Curves of Chemically Modified Wheat Straw Filler-Based Biocomposites Using Machine Learning Techniques. *Journal of Composites Science, 9*(5), 221.

Hazmi, B., Farooq, H., Rashid, U., Ghani, W. A. W. A. K., Yaw, T. C. S., Ngamcharussrivichai, C., & Ali, I. (2026). Characterization and pyrolysis kinetic modelling of lignocellulosic waste from rambutan seeds: A machine learning approach. *Biomass and Bioenergy, 204*, 108426.

Kartal, F., Dalbudak, Y., & Özveren, U. (2023). Prediction of thermal degradation of biopolymers in biomass under pyrolysis atmosphere by means of machine learning. *Renewable Energy, 204*, 774-787.

Khan, H., Savvopoulos, S., & Janajreh, I. (2024). Artificial neural network-assisted thermogravimetric analysis of thermal degradation in combustion reactions: A study across diverse organic samples. *Environmental Research, 249*, 118463.

Kim, H., Jo, H., & Ryu, C. (2024). Derivation of kinetic parameters and lignocellulosic composition from thermogram of biomass pyrolysis using convolutional neural network. *International Journal of Energy Research, 2024*(1), 6184508.

Mohammadpour, A., Dolatabadi, M., Bontempi, E., & Shahsavani, E. (2025). Synthesis and characterization of novel lignocellulosic biomass-derived activated carbon for dye removal: Machine learning optimization, mechanisms, and antibacterial properties. *Biomass and Bioenergy, 192*, 107490.

Otaru, A. J., & Albin Zaid, Z. A. A. (2025). Analysis of TGA data for polyvinyl alcohol at slow heating rate using deep neural networks, activation energy, and activation enthalpy. *Scientific Reports, 15*(1), 37915.

Otaru, A. J., Albin Zaid, Z. A. A., Alkhaldi, M. M., Albin Zaid, S. M. A., & AlShuaibi, A. (2025). The Bioenergy Potential of Date Palm Branch/Waste Through Reaction Modeling, Thermokinetic Data, Machine Learning KNN Analysis, and Techno-Economic Assessments (TEA). *Polymers, 17*(23), 3182.

Pambudi, S., Jongyingcharoen, J. S., & Saechua, W. (2025). Explainable machine learning for predicting thermogravimetric analysis of oxidatively torrefied spent coffee grounds combustion. *Energy, 320*, 135288.

Park, M., Um, B. H., Park, S.-H., & Kim, D.-Y. (2025). Exploring the Feasibility of Deep Learning for Predicting Lignin GC-MS Analysis Results Using TGA and FT-IR. *Polymers, 17*(6), 806.

Velázquez-Martí, B., Bonini-Neto, A., Leão-dos-Santos, W. P., Gaibor-Chávez, J., Escobar-Machado, J. A., & Álvarez-Montero, X. (2025). Using artificial neural networks for classification of composition and biomass species for energy based on thermogravimetric data. *International Journal of Energy Research, 2025*(1), 8832502.

Xiao, K., & Zhu, X. (2024). Machine learning approach for the prediction of biomass waste pyrolysis kinetics from preliminary analysis. *ACS omega, 9*(49), 48125-48136.

Yao, Y., Wei, G., Yuan, H., Kang, Z., Huang, Z., Yang, X., . . . Xie, J. (2025). Atmosphere-driven mechanisms in biogas residue chemical looping pyrolysis: Insights from kinetic characteristic and machine learning prediction. *Fuel, 391*, 134691.

Yin, X., Tao, J., Wang, J., Yan, B., Chen, G., & Cheng, Z. (2025). Prediction of activation energy of lignocellulosic biomass pyrolysis through thermogravimetry-assisted machine learning. *Biomass and Bioenergy, 194*, 107644.

Zaifullizan, Y. M., Kuan, L. M., Salema, A. A., & Ishaque, K. (2023). Comparison of artificial intelligence models to predict oil palm biomass pyrolysis and kinetics using thermogravimetric analysis. *Journal of Oil Palm Research, 35*(1), 86-99.

Zhong, Y., Liu, F., Huang, G., Zhang, J., Li, C., & Ding, Y. (2024). Thermogravimetric experiments based prediction of biomass pyrolysis behavior: A comparison of typical machine learning regression models in Scikit-learn. *Marine Pollution Bulletin, 202*, 116361.