UDC 536.6
IRSTI 29.19.09

https://doi.org/10.55452/1998-6688-2026-23-3-398-413

[1*]Iztleuova B.,
PhD student, ORCID ID: 0009-0009-7588-8976,
[*]e-mail: b.iztleuova@zhubanov.edu.kz
[1]Imanchiyev A.,
Associate Professor, candidate of physical and mathematical sciences,
ORCID ID: 0000-0002-1835-2501,
e-mail: imanchiev_ae@mail.ru
[1]Bekeshev A.,
Associate Professor, candidate of physical and mathematical sciences,
ORCID ID: 0000-0002-7038-4631
e-mail: amirbek2401@gmail.com
[1]Zhanturina N.,
Associate Professor, PhD, ORCID ID: 0000-0001-9540-6334,
e-mail: nzhanturina@mail.ru
[2]Ussenkulova Sh.,
Associate Professor, PhD, ORCID ID: 0000-0001-9417-755X,
e-mail: Sholpan1990@gmail.com
[3]Wang X.,
Associate Professor, PhD, ORCID ID: 0000-0001-5881-4400,
e-mail: wxcmx@ustc.edu.cn

[1]K. Zhubanov Aktobe Regional University, Aktobe, Kazakhstan
[2]Kazakh University of Technology and Business, Astana, Kazakhstan
[3]State Key Laboratory of Fire Science, University of Science and Technology of China,
Hefei, China

EMPIRICAL BENCHMARKING OF GRADIENT BOOSTING FOR THERMAL DEGRADATION PREDICTION IN MINERAL–POLYMER COMPOSITES

Abstract

This study presents a mathematical and empirical analysis of gradient boosting algorithms applied to thermogravimetric analysis (TGA) data for composites based on vinyl ether matrices and epoxides with mineral fillers. The gradient boosting algorithm is output step by step with a numerical example of TGA. XGBoost and CatBoost are then analyzed formally. 12 algorithms were compared on 478 data from the peer-reviewed literature using 5-repeated 5-fold cross-validation. Extra Trees and CatBoost achieved statistically equivalent top performance (mean $R^2 = 0.7430 \pm 0.0879$ and 0.7313 ± 0.0793, respectively; Nadeau & Bengio corrected $t = 0.483$, $p = 0.634$), outperforming all linear baselines, SVR, and MLP. CatBoost is recommended for practical deployment: its ordered boosting produces unbiased gradient estimates on small datasets ($n < 1000$), and SHAP analysis of its gradient tree structure reveals physically interpretable feature attributions aligned with known TGA mechanisms. SHAP values identified TGA peak temperature (T_{max}) and matrix type as the dominant predictors. The proposed preprocessing pipeline including phr to wt% conversion, label encoding, and median imputation provides a reproducible strategy for heterogeneous TGA literature datasets.

Keywords: gradient boosting, XGBoost, CatBoost, polymer composites, machine learning, thermogravimetric analysis, ordered boosting, mathematical foundations, algorithm selection.

Received April 15, 2026; revised July 7, 2026; accepted July 17, 2026

Introduction

Machine learning has reshaped how researchers address property prediction in materials science and polymer engineering. Characterizing composite systems experimentally is both costly and time-consuming, and the composition–property space is far too broad to map by trial and error. Machine learning methods are now common in materials science. Researchers use them to find links between how a material is built and how it behaves. No physical model is needed for that [1, 2, 3]. When the data are in tabular form, gradient boosting works well. It picks up nonlinear patterns even when sample sizes are small or data are messy. That is why materials scientists often reach for it [4, 5, 6]. The method works through shallow decision trees built one after another. Every new tree looks at what the previous one got wrong and tries to fix it. The loss function decreases with each step [7]. Two variants stand out in polymer and composite research. XGBoost keeps trees from becoming too large by using a second-order Taylor expansion and adding L1 and L2 penalty terms to the loss [8]. CatBoost uses permutation-based gradient estimation, which reduces target leakage compared to standard gradient boosting [9].

Between 2018 and 2024 alone, more than 170 publications applied boosting methods in polymer science. Both algorithms now appear routinely, from thermal property prediction to mechanical strength modeling [1, 6, 10]. However, here is the problem: across nearly all of that work, the focus remains on predictive accuracy [11, 12, 13, 14]. Why XGBoost regularizes the way it does, or how exactly CatBoost avoids overfitting on small data, these questions rarely receive a serious answer in the polymer composite literature.

The aim of this study is to benchmark gradient boosting against a broad range of regression models and to establish its standing for predicting the thermal degradation temperature (T50%) of mineral–polymer composites, justifying the recommended model on mathematical and SHAP-based grounds. This paper makes the following contributions. First, we provide a unified step-by-step mathematical derivation of gradient boosting, XGBoost, and CatBoost applied to TGA data of mineral–polymer composites. Second, we benchmark twelve regression algorithms on 478 TGA samples using 5-repeated 5-fold cross-validation, providing statistically reliable performance estimates. Third, we interpret the best-performing model using SHAP analysis, linking feature contributions to the physical mechanisms of thermal degradation. While boosting methods have been widely applied in materials science [1, 6, 10], their mathematical foundations, rigorous cross-validated benchmarking, and SHAP-based interpretation on TGA composite data remain underexplored. Empirically, gradient boosting is confirmed among the top performers and is statistically indistinguishable from the best model; the contribution of this work is therefore methodological rather than algorithmic, providing a principled, SHAP-supported model recommendation on a reproducible pipeline for heterogeneous, literature-aggregated TGA data.

Materials and methods

Gradient Boosting: Mathematical Framework

Gradient boosting is used to construct a prediction model as a sum of M weak learners:

$$\hat{y}_i = F_M(x_i) = \sum_{m=1}^{M} f_m(x_i) \quad (1)$$

where each f_m is a shallow decision tree and x_i is the feature vector of the i-th sample. The model is built sequentially at each step m, and a new tree f_m is added to correct the residual error of the current ensemble:

$$F_m(x) = F_{m-1}(x) + \eta \cdot f_m(x) \quad (2)$$

where η∈(0,1] is the learning rate, which controls the contribution of each new tree. Small values of η slow down learning but improve generalization by preventing any single tree from dominating the ensemble [7].

To illustrate the gradient boosting algorithm (Table 1), we apply it to a small dataset derived from thermogravimetric analysis (TGA) of vinyl, epoxy composites. The goal is to predict the maximum decomposition temperature T_max based on two input features: mineral loading (%) and modifier loading (%).

Table 1 – Illustrations of the TGA dataset of epoxy composites used for gradient boosting

№	Loading of mineral	Loading of modifier	tga_tmax
1	0	0	390
2	4.76	0.05	386
3	9	0.09	389
4	30	5	373
5	30	5	379

The algorithm begins with a constant prediction F_0 equal to the mean of all target values:

$$F_0 = \frac{390 + 386 + 389 + 373 + 379}{5} = 383.4\,°C \quad (3)$$

The residuals $r_i=y_i-F_0$ represent the errors of the current model. For example, sample 1 yields $r_1=390-383.4=+6.6$, while sample 4 yields $r_4=373-383.4=-10$.

Table 2 – Initial residuals computed as $r_i=y_i-F_0$

№	y	F_0	Residual r
1	390	383.4	+6.6
2	386	383.4	+2.6
3	389	383.4	+5.6
4	373	383.4	-10.4
5	379	383.4	-4.4

At each node of the decision tree, the algorithm evaluates every possible split across all features and all thresholds. For each candidate split, the data are divided into two groups, left and right, and the Mean Squared Error (MSE) is computed for each group separately [15]. The MSE for a group is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})^2 \quad (4)$$

where r_i is the residual of sample i, $\bar{r}$ is the mean residual of the group and n is the number of samples in the group. The total MSE of a split is the weighted sum of both groups:

$$MSE_{total} = \frac{n_L}{n} \cdot MSE_L + \frac{n_R}{n} MSE_R \quad (5)$$

where n_L and n_R are the numbers of samples in the left and right groups, respectively. For example, when the split mineral loading > 9 on the residuals from Table 1 is evaluated:

Left group (samples 1, 2, 3): residuals = 6.6, 2.6, and 5.6, respectively.

$$\bar{r_L} = \frac{6.6 + 2.6 + 5.6}{3} = 4.93 \quad (6)$$

$$MSE_L = \frac{(6.6 - 4.93)^2 + (2.6 - 4.93)^2 + (5.6 - 4.93)^2}{3} = 2.89 \quad (7)$$

Right group (samples 4, 5): residuals = −10.4, −4.4.

$$\bar{r_R} = \frac{-10.4 + (-4.4)}{2} = -7.4 \quad (8)$$

$$MSE_R = \frac{(-10.4 + 7.4)^2 + (-4.4 + 7.4)^2}{2} = 9.0 \quad (9)$$

Total MSE:

$$MSE_{total} = \frac{3}{5} \cdot 2.89 + \frac{2}{5} \cdot 9.0 = 1.73 + 3.6 = 5.33 \quad (10)$$

Split: mineral loading > 4.76. Left group (samples 1 and 2): residuals = 6.6 and 2.6.

$$\bar{r_L} = \frac{6.6 + 2.6}{2} = 4.6 \quad (11)$$

$$MSE_L = \frac{(6.6 - 4.6)^2 + (2.6 - 4.6)^2}{2} = 4.0 \quad (12)$$

Right group (samples 3, 4, and 5): residuals = 5.6, −10.4, and −4.4.

$$\bar{r_R} = \frac{5.6 + (-10.4) + (-4.4)}{3} = -3.07 \quad (13)$$

$$MSE_R = \frac{(5.6 + 3.07)^2 + (-10.4 + 3.07)^2 + (-4.4 + 3.07)^2}{3} = 43.56 \quad (14)$$

Total MSE:

$$MSE_{total} = \frac{2}{5} \cdot 4.0 + \frac{3}{5} \cdot 43.56 = 1.6 + 26.13 = 27.73 \quad (15)$$

All the candidate splits are compared in the same manner, and the split with the lowest total MSE is selected as the best node. In this example, a split mineral loading > 9 produced the lowest MSE among all the evaluated thresholds and features and was therefore chosen as the root node of the decision tree. The tree evaluates all the features and thresholds. Best split: loading_mineral > 9:

Loading_mineral > 9?
/ \
No Yes
samples 1,2,3 samples 4,5
mean = +4.93 mean = -7.4

A shallow decision tree is trained on the residuals. The optimal split is found at mineral loading > 9%, with two leaf nodes having mean values of +4.93 and −7.4. The model is updated using a learning rate η=0.1:

$$F_1(x) = F_0(x) + \eta \cdot f_1(x) \tag{16}$$

Table 3 – Updated model predictions F_1 and new residuals after the first iteration

№	y	F_1	New residual
1	390	383.4 + 0.1×4.93 = 383.9	6.1
2	386	383.4 + 0.1×4.93 = 383.9	2.1
3	389	383.4 + 0.1×4.93 = 383.9	5.1
4	373	383.4 + 0.1×(−7.4) = 382.7	−9.7
5	379	383.4 + 0.1×(−7.4) = 382.7	−3.7

This process is repeated M times. With each iteration, the residuals decrease, and the predictions become more accurate. The overall objective to minimize combines a loss term and a regularization term:

$$\mathcal{L} = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{m=1}^{M} \Omega f_m \tag{17}$$

where $l(\hat{y}_i, y_i)$ is the discrepancy between the predicted and actual values and Ωf_m is the penalized tree complexity. The exact form of Ω differs between XGBoost and CatBoost and is discussed in Sections 3 and 4, respectively [8]. Despite this growing adoption, the mathematical foundations behind these implementations remain largely absent from the polymer and composite literature.

XGBoost: Mathematical Foundations

All mathematical formulations presented in this section are derived from Chen & Guestrin [8]. Unlike classical gradient boosting, which minimizes the loss function approximately using only first-order gradients, XGBoost employs a second-order Taylor expansion of the loss function at each iteration:

$$\mathcal{L}^{(m)} \approx \sum_{i=1}^{n} \left[l\left(y_i, F_{m-1}(x_i)\right) + g_i f_m(x_i) + \frac{1}{2} h_i f_m^2(x_i) \right] \tag{18}$$

where g_i and h_i are the first- and second-order derivatives of the loss function with respect to the previous prediction $F_{m-1}(x_i)$:

$$g_i = \frac{\partial l(y_i, F_{m-1})}{\partial F_{m-1}}, \quad h_i = \frac{\partial^2 l(y_i, F_{m-1})}{\partial F_{m-1}^2} \tag{19}$$

Compared with the first-order methods, the Hessian term h_i captures the curvature of the loss surface, providing the algorithm with a more accurate picture of how the loss changes. These processes result in more accurate step estimation than classical gradient boosting, which relies solely on the first-order gradient.

To control model complexity and prevent overfitting, XGBoost incorporates an explicit regularization term $\Omega(f_m)$ directly into the objective function:

$$\mathcal{L}^{(m)} \approx \sum_{i=1}^{n} \left[g_i f_m(x_i) + \frac{1}{2} h_i f_m^2(x_i) \right] + \Omega(f_m) \tag{20}$$

where the regularization term is defined as follows:

$$\Omega(f_m) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 + \alpha \sum_{j=1}^{T} |\omega_j| \tag{21}$$

where T is the number of leaves in the tree; ω_j is the weight of leaf j; γ is the penalty for the number of leaves (tree complexity); λ is the L2 regularization coefficient (ridge); and α is the L1 regularization coefficient (lasso).

The L2 term penalizes large leaf weights, whereas the L1 term encourages sparsity by pushing small weights to zero. By grouping samples that fall into the same leaf j, the objective function can be rewritten as a sum over leaves:

$$\mathcal{L}^{(m)} \approx \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) \omega_j(x_i) + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma \mathrm{T} \tag{22}$$

Since this is a quadratic function with respect to ω_j, the optimal weight for each leaf is obtained analytically by setting the derivative to zero:

$$\frac{\partial \mathcal{L}}{\partial \omega_j} = \sum_{i \in I_j} g_i + \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j = 0 \tag{23}$$

Solving for ω_j:

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{24}$$

Substituting ω_j back into the objective yields the minimum loss value for a given tree structure:

$$\mathcal{L}^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{25}$$

This closed-form solution allows XGBoost to evaluate the quality of any tree structure exactly, without numerical optimization.

CatBoost Mathematical Distinctions

All mathematical formulations presented in this section are derived from Prokhorenkova et al. [9]. In classical gradient boosting, the residuals for all training samples are computed using a model that is trained on the same samples. Formally, at iteration m, the gradient g_i for sample i is computed as follows:

$$g_i = \frac{\partial l(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \tag{26}$$

where F_{m-1} was built using y_i itself. This creates target leakage, and the target value y_i influences both the model and the gradient estimate for the same observation, leading to biased gradient estimation and overfitting, which are particularly critical for small datasets.

CatBoost addresses this through ordered boosting, where each gradient is computed from a model that has only seen samples preceding the current one in a randomly assigned order σ:

$$g_i^{\sigma} = \frac{\partial l(y_i, F_m^{\sigma(i)}(x_i))}{\partial F_m^{\sigma(i)}(x_i)} \quad (27)$$

where $F_m^{\sigma(i)}$ denotes a model built only on samples $\{x_j : \sigma(j) < \sigma(i)\}$.

This ensures that when the gradient for sample i is computed, its target value y_i has never been seen by the model, producing an unbiased gradient estimate:

$$\mathbb{E}[g_i^{\sigma}] = \mathbb{E}\left[\frac{\partial l(y_i, F_{m-1})}{\partial F_{m-1}}\right] \quad (28)$$

To implement this, CatBoost keeps n model versions running in parallel, each updated as new samples enter the permutation. The greatest advantage of ordered boosting is precisely where classical GB struggles most: small datasets with high exposure to target leakage.

Dataset, Feature Engineering and Preprocessing

The dataset was compiled from peer-reviewed publications reporting TGA results for vinyl ether and epoxy matrices with mineral fillers. The prediction target for the machine learning benchmark is T50% (the temperature at which 50% mass loss occurs during TGA), extracted from each TGA curve at tga_x_pct = 50. Note that T_max (the DTG peak temperature) is used as a predictor input feature, not as the target variable; it served as the illustrative target only in the worked example of Section 2. Table 4 shows the features used and the preprocessing applied before model training.

Table 4 – Features, preprocessing characteristics before training models

Features	Unit of measure	Preprocessing
matrix type	VE/EP	Vinyl ester, epoxy
mineral formula	Chemical formula	-
mineral loading	wt%	Phr converted to wt%
TGA peak temperature (T_max)	°C	.→, 4.5→4,5
mineral name	-	Label encoded (categorical)
residue measurement temperature (tga_ residue_t)	°C	Median imputation
TGA residue mass fraction (tga_residue)	wt%	Median imputation

Missing numerical values were imputed with the column median. The TGA atmosphere (N_2, air, O_2, Ar, CO_2) was examined but found to be uninformative for this dataset: 99% of records were measured under nitrogen. This feature was therefore excluded from the final model.

Cross-Validation Strategy

All performance metrics reported in Table 5 and the scatter plot in Figure 1 are derived from the same 5-repeated 5-fold cross-validation (RepeatedKFold, n_splits = 5, n_repeats = 5, random_ state = 42). Figure 1 displays out-of-fold (OOF) predictions: each of the 478 samples is predicted exactly once per repeat by a model that has not observed it, so no separate train/test split is required for visualisation. Model performance was estimated using 5-repeated 5-fold cross-validation (RepeatedKFold, n_repeats = 5, n_splits = 5, random_state = 42), yielding 25 independent evaluations per model across the full 478-sample dataset.

Models and Evaluation Metrics

12 regression algorithms were trained and evaluated: Linear Regression, Ridge Regression (α = 1.0), Huber Regression (ε = 1.35), Support Vector Regression (SVR, C = 100, ε = 5), Multilayer Perceptron (MLP; 128–64–32 neurons), Random Forest (300 trees), Extra Trees (300 trees), Gradient

Boosting (400 estimators, lr = 0.05, depth = 4, subsample = 0.8), Gradient Boosting with Huber loss (α = 0.9), XGBoost (400 estimators, lr = 0.05, depth = 5, subsample = 0.8, colsample_bytree = 0.8), CatBoost (500 iterations, lr = 0.05, depth = 6, l2_leaf_reg = 3), and a Stacking ensemble combining Gradient Boosting, XGBoost, and CatBoost with a Ridge meta-learner (5-fold cross-validation). Hyperparameter values were fixed a priori according to established recommendations in the machine learning literature [8, 9, 16, 17] rather than tuned per model; no grid search or nested cross-validation was performed, so every algorithm was evaluated under comparable, untuned conditions. The chosen values were confirmed to be stable across the 25 cross-validation evaluations.

All metrics were computed across 25 CV folds and reported as mean ± std. All error metrics are expressed as percentages to facilitate cross-study comparison:

- CV(RMSE) %: coefficient of variation of RMSE normalizes prediction error to the scale of the target variable;
- MAPE %: mean absolute percentage error expresses average error as a percentage of each individual prediction;
- MAE %: relative mean absolute error percentage counterpart of the absolute MAE;
- R^2 (mean ± std): coefficient of determination averaged over 25 folds with standard deviation fraction of T50% variance explained; std quantifies model stability.

Results

Model Comparison

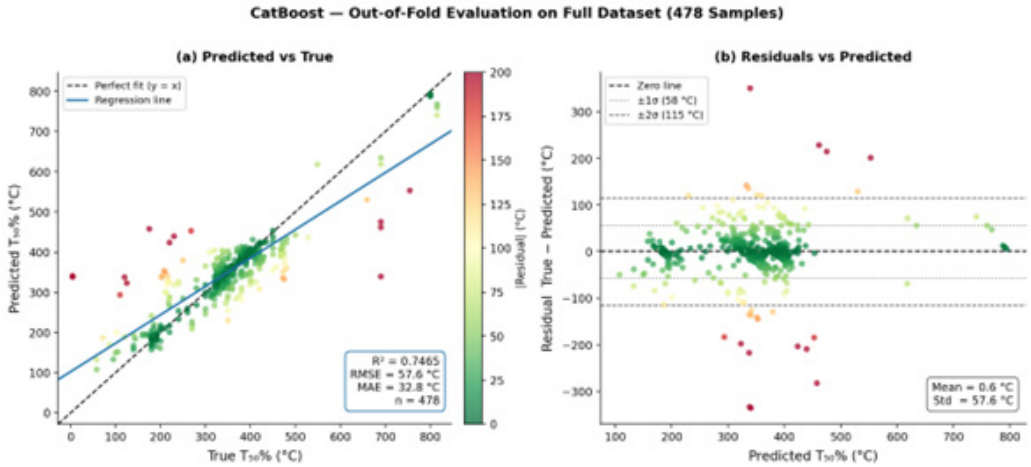
Table 5 summarizes the performance of all 12 models evaluated by 5-repeated 5-fold cross-validation (25 evaluations per model). The best overall performance was achieved by Extra Trees, with mean R^2 = 0.7430 ± 0.0879, CV(RMSE) = 15.89%, MAPE = 10.26%, and MAE = 8.72%. Extra Trees and CatBoost achieved comparable top performance (mean R^2 = 0.7430 ± 0.0879 vs. 0.7313 ± 0.0793). A Nadeau & Bengio corrected resampled t-test [18] – which accounts for the non-independence of folds in repeated k-fold CV – yielded t = 0.483 (df = 24, p = 0.634, two-tailed), confirming that the difference between the two models is not statistically significant. All tree-based ensemble methods (Extra Trees, CatBoost, XGBoost, Gradient Boosting) consistently outperformed linear baselines and SVR, confirming the non-linear nature of the T50% response surface.

Table 5 – Performance metrics of all regression models evaluated by 5-repeated 5-fold cross-validation (478 samples, 25 evaluations per model). Models are ranked by mean R^2 (descending)

Model	mean R^2 ± std	CV(RMSE) %	MAPE %	MAE %
Extra Trees	0.7430 ± 0.0879	15.89	10.26	8.72
CatBoost	0.7313 ± 0.0793	16.37	10.4	9.5
XGBoost	0.7256 ± 0.0805	16.48	10.94	9.71
GB (Huber loss)	0.7078 ± 0.1036	16.97	9.81	9.11
Gradient Boosting	0.7060 ± 0.0897	17.06	10.75	9.79
Random Forest	0.6963 ± 0.0950	17.33	10.97	9.89
Stacking (GB+XGB+CAT)	0.6633 ± 0.0825	18.46	12.13	11.23
SVR	0.6432 ± 0.0964	18.95	10.88	10.4
MLP	0.5536 ± 0.1165	21.09	15.86	14.36
Ridge	0.4891 ± 0.1420	22.56	15.6	14.45
Linear Regression	0.4889 ± 0.1425	22.56	15.6	14.43
Huber Regression	0.4529 ± 0.1946	23.13	17.93	13.57

Predicted vs. True T50% and Residual Analysis

Figure 1 shows out-of-fold scatter plots for the CatBoost model: predicted versus true T50% values (left panel) and the corresponding residual distribution (right panel), evaluated by 5-repeated 5-fold cross-validation across all 478 samples. A small number of samples (n < 5) show residuals exceeding 150 °C, attributable to inter-source measurement variability inherent in a dataset aggregated from multiple independent publications.

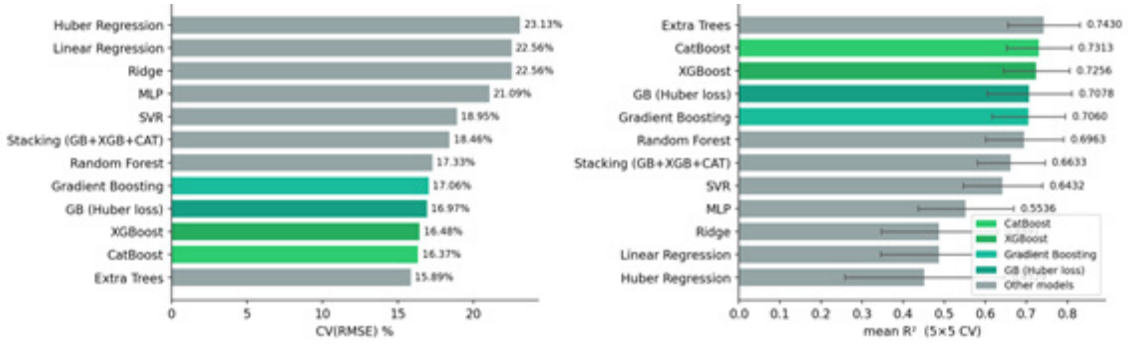


Predicted vs. True T50% (top row) and residual plots (bottom row) for the three best models. The dashed line represents the ideal y = x relationship. Dotted lines indicate ±1σ residual bounds

Figure 1– CatBoost out-of-fold predictions (5-repeated 5-fold cross-validation, n = 478):
(a) predicted vs. true T_{50}% values; (b) residuals vs. predicted T_{50}%.
Colour encodes absolute residual magnitude

Model Ranking: RMSE and R^2

Figure 2 presents horizontal bar charts ranking all models by RMSE (lower is better) and R^2 (higher is better). Gradient boosting models (highlighted in green) consistently ranked at the top, while linear models ranked at the bottom.

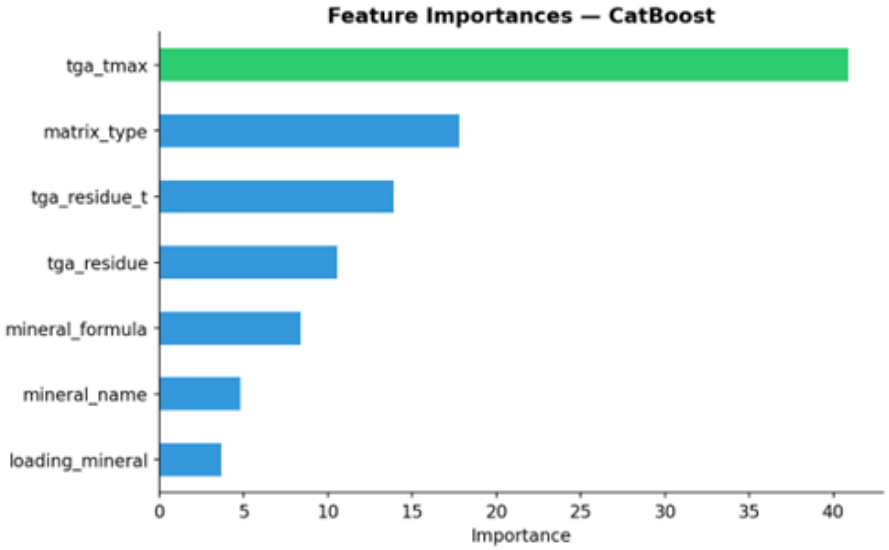


Green bars: Extra Trees, CatBoost, XGBoost, Gradient Boosting. Grey bars: remaining algorithms. Error bars = ±1 std.

Figure 2 – CV(RMSE) % (left) and mean R^2 ± std (right)
for all evaluated models under 5-repeated 5-fold cross-validation

SHAP Analysis

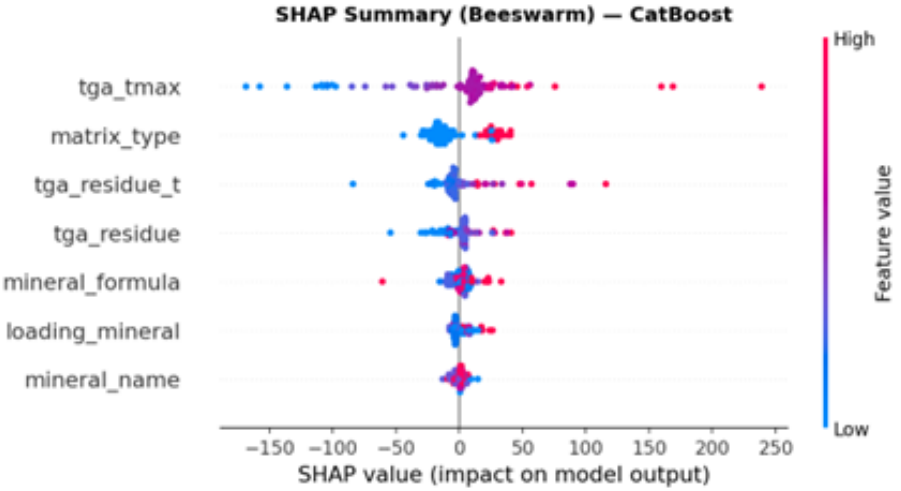
Figure 3 displays the feature importances derived from the CatBoost model.



The highlighted bar indicates the most important feature.

Figure 3 – Feature importance for the CatBoost model

Figure 4 shows the SHAP summary plot for CatBoost.



Left: beeswarm summary plot showing the distribution of SHAP values for all test samples across seven input features; each point represents one sample, color indicates the feature value (pink – high, blue – low).

Figure 4 – SHAP analysis of the CatBoost model

Figure 5 illustrates SHAP dependence plots for the top features.

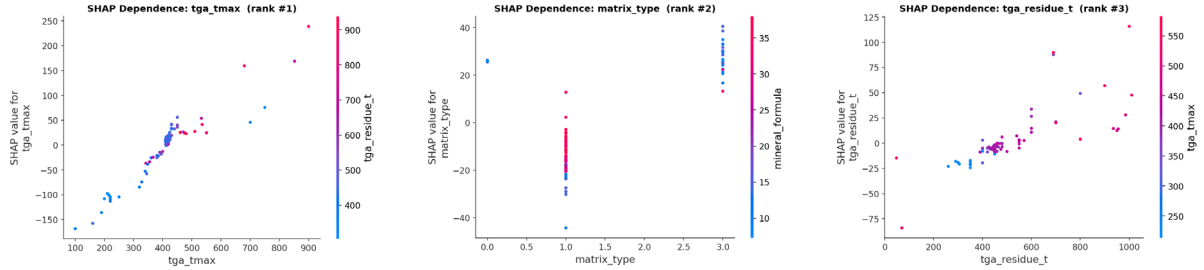


Figure 5-SHAP dependence plots for the three most influential features in the CatBoost model: tga_tmax (rank 1), matrix_type (rank 2), and tga_residue_t (rank 3).

Each point represents one test sample; the y-axis shows the SHAP contribution of the feature, and the color indicates the value of the most interacting variable.

Discussion

Tree-based ensemble methods consistently outperformed linear regression, Ridge, and SVR across all cross-validation folds. This confirms that the composition–degradation relationship in mineral–polymer composites is inherently non-linear [10, 27]. Filler type, matrix chemistry, and loading interact in ways that linear models cannot capture [1, 12]. Extra Trees and CatBoost achieved comparable top performance (mean $R^2 = 0.7430 \pm 0.0879$ vs. 0.7313 ± 0.0793). A Nadeau & Bengio corrected resampled t-test confirmed that this difference is not statistically significant ($t = 0.483$, $df = 24$, $p = 0.634$). Similar parity between these algorithms has been reported on other heterogeneous materials datasets [29,32]. Despite the marginal numerical edge of Extra Trees ($\Delta R^2 = 0.012$, not statistically significant per the Nadeau & Bengio corrected test), CatBoost is recommended for practical deployment on the basis of two complementary properties. First, ordered boosting (Eq. 27–28) eliminates the target-leakage bias inherent in both classical gradient boosting and XGBoost when sample counts are low; on a dataset of $n = 478$ heterogeneous TGA samples, bias reduction provides a more reliable generalisation guarantee than the variance reduction achieved by Extra Trees through extreme randomisation [9, 17]. Second, SHAP analysis (Figures 4–5) corroborates CatBoost's practical value: feature attributions align with established physical mechanisms, including the thermal coupling encoded in T_max, the polymer chemistry signal in matrix_type, and the non-linear saturation behaviour of mineral loading, yielding a mechanistically interpretable model [30, 31, 32, 33]. This produces more stable predictions on unseen mineral–polymer combinations outside the training distribution [17, 19]. On small heterogeneous datasets such as this one ($n = 478$), this property is particularly valuable [13, 17].

XGBoost ranked third, with mean $R^2 = 0.7256 \pm 0.0805$, closely behind CatBoost (0.7313 ± 0.0793). The difference between the two is small and not statistically significant. This is consistent with prior observations that XGBoost's L1/L2 regularization advantage is most pronounced on large clean datasets [8, 16]. On a small heterogeneous dataset ($n = 478$), ordered boosting in CatBoost reduces target leakage more effectively than XGBoost's regularization, giving CatBoost a marginal edge [9, 17].

The dataset aggregates TGA measurements from multiple laboratories and instruments. This introduces inter-source variability in decomposition temperatures that a single-site study would avoid [34, 13]. Heating rates, sample mass, and purge gas flow all vary across publications and affect measured T_{50}% values [34].Label encoding assigns arbitrary integers to polymer matrix types and mineral names. This representation does not capture chemical similarity between structurally related matrices or minerals [35, 11]. Graph-based or SMILES-derived descriptors may improve generalization to unseen polymer–filler combinations [35, 11]. The current model predicts T_{50}% only. Extending the framework to T_{10}%, T_max, and char yield would provide a more complete description of composite thermal behavior [1, 4]. This remains future work.

SHAP analysis of the CatBoost model (Figures 4–5) provided mechanistically interpretable feature rankings [28,33, 36]. T_max dominated predictions across all test samples. Its high positive SHAP values reflect the physical coupling between the DTG peak temperature and the mass-loss midpoint along the TGA curve [37, 38]. Matrix type ranked second, reflecting the fundamental role of polymer chemistry in governing thermal decomposition behavior [1, 12]. The residue measurement temperature (tga_residue_t) ranked third, acting as a proxy for the high-temperature thermal stability plateau [37, 39]. Mineral loading showed a non-linear SHAP profile. Moderate filler contents contributed positively to predicted T50%, while extreme loadings reduced model confidence, consistent with known saturation effects in filler-reinforced composites [4, 5, 40]. The remaining categorical descriptors (mineral formula, mineral name) contributed collectively but ranked below the top three features. This suggests that the dominant thermal response is governed by measurable thermophysical parameters rather than compositional class labels [38, 41].

CatBoost's ordered boosting algorithm (Eq. 27–28) eliminates the target leakage present in both classical GB and XGBoost, producing unbiased gradient estimates that are particularly valuable when $n < 1000$ [9, 17]. This explains why CatBoost outperforms XGBoost despite the latter's more sophisticated regularization: on a heterogeneous 478-sample dataset, bias reduction through ordered boosting provides a greater accuracy benefit than L1/L2 penalty control [8, 16]. The underperformance of GB(Huber loss) relative to tree-ensemble methods (Extra Trees, Random Forest) further suggests that the dataset's predictive signal is better captured through variance reduction via bagging than through boosting's sequential residual correction [15]. XGBoost ranked third, consistent with prior benchmarks showing that its regularization advantage is most pronounced on clean, large-scale structured datasets [16, 20, 21]. CatBoost tends to outperform XGBoost on small or noisy samples, as demonstrated in molecular property prediction tasks [17, 24] and domain-specific comparisons [22, 23]. On small tabular datasets, tree-based ensemble methods generally outperform deep learning approaches, and boosted trees remain the preferred choice when sample counts are low [25, 26].

Conclusion

T_max dominated predictions across all test samples: its high positive SHAP values reflect the physical coupling between the DTG peak temperature and the mass-loss midpoint along the TGA curve [37, 38]. Matrix type ranked second, reflecting the fundamental role of polymer chemistry in governing thermal decomposition behavior [1, 12]. The residue measurement temperature (tga_residue_t) ranked third, acting as a proxy for the high-temperature thermal stability plateau [37, 39]. Mineral loading showed a non-linear SHAP profile: moderate filler contents contributed positively to predicted T_{50}%, while extreme loadings reduced model confidence, consistent with known saturation effects in filler-reinforced composites [4, 5]. Categorical descriptors (mineral formula, mineral name) contributed collectively but ranked below continuous TGA-derived variables, suggesting that the dominant thermal response is governed by measurable thermophysical parameters rather than compositional class labels [38, 41].

Funding. This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant No. BR28712729.

REFERENCES

1 Karuppusamy, M., Thirumalaisamy, R., Palanisamy, S., Nagamalai, S., El Sayed Massoud, E., and Ayrilmis, N. A review of machine learning applications in polymer composites: advancements, challenges, and future prospects. Journal of Materials Chemistry A, 13, 16290–16308 (2025). https://doi.org/10.1039/D5TA00982K

2 Martin, T.B., and Audus, D.J. Emerging trends in machine learning: a polymer perspective. ACS Polymers Au, 3, 239–258 (2023). https://doi.org/10.1021/acspolymersau.2c00053

3 Pai, S.M., Shah, K.A., Sunder, S., Albuquerque, R.Q., Brütting, C., and Ruckdäschel, H. Machine learning applied to the design and optimization of polymeric materials: a review. Next Materials, 7, 100449 (2025). https://doi.org/10.1016/j.nxmate.2024.100449

4 Jafari, P. et al. Machine learning for expediting next-generation of fire-retardant polymer composites. Composites Communications, 45, 101806 (2024). https://doi.org/10.1016/j.coco.2023.101806

5 Xiao, J., Hobson, J., Ghosh, A., Haranczyk, M., and Wang, D.-Y. Flame retardant properties of metal hydroxide-based polymer composites: a machine learning approach. Composites Communications, 40, 101593 (2023). https://doi.org/10.1016/j.coco.2023.101593

6 Berladir, K., Antosz, K., Ivanov, V., and Mitaľová, Z. Machine learning-driven prediction of composite materials properties based on experimental testing data. Polymers, 17, 694 (2025). https://doi.org/10.3390/polym17050694

7 Friedman, J.H. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29, 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451

8 Chen, T., and Guestrin, C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

9 Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., and Gulin, A. CatBoost: unbiased boosting with categorical features. arXiv:1706.09516 (2019). https://doi.org/10.48550/arXiv.1706.09516

10 Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V., and Borodulin, A. Boosting-based machine learning applications in polymer science: a review. Polymers, 17, 499 (2025). https://doi.org/10.3390/polym17040499

11 Ge, W., De Silva, R., Fan, Y., Sisson, S.A., and Stenzel, M.H. Machine learning in polymer research. Advanced Materials, 37, 2413695 (2025). https://doi.org/10.1002/adma.202413695

12 Han, J. et al. Toward accurate machine learning-driven prediction of polymeric composites properties based on experimental data. Materials Genome Engineering Advances, 3, e70027 (2025). https://doi.org/10.1002/mgea.70027

13 Xu, P., Ji, X., Li, M., and Lu, W. Small data machine learning in materials science. npj Computational Materials, 9, 42 (2023). https://doi.org/10.1038/s41524-023-01000-z

14 Zhong, X., Gallagher, B., Liu, S., Kailkhura, B., Hiszpanski, A., and Han, T.Y.-J. Explainable machine learning in materials science. npj Computational Materials, 8, 204 (2022). https://doi.org/10.1038/s41524-022-00884-7

15 Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. Classification and Regression Trees. Routledge, New York (1984). https://doi.org/10.1201/9781315139470

16 Florek, P., and Zagdański, A. Benchmarking state-of-the-art gradient boosting algorithms for classification. arXiv:2305.17094 (2023). https://doi.org/10.48550/arXiv.2305.17094

17 Boldini, D., Grisoni, F., Kuhn, D., Friedrich, L., and Sieber, S.A. Practical guidelines for the use of gradient boosting for molecular property prediction. Journal of Cheminformatics, 15, 73 (2023). https://doi.org/10.1186/s13321-023-00743-7

18 Nadeau, C., and Bengio, Y. Inference for the generalization error. Machine Learning, 52, 239–281 (2003). https://doi.org/10.1023/A:1024068626366

19 Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937–1967 (2021). https://doi.org/10.1007/s10462-020-09896-5

20 Cakiroglu, M., Bekdaş, G., and Geem, Z.W. Fatigue predictive modeling of composite materials for wind turbine blades using explainable gradient boosting models. Coatings, 15, 325 (2025). https://doi.org/10.3390/coatings15030325

21 Mustapha, I.B., Abdulkareem, M., Jassam, T.M., AlAteah, A.H., Al-Sodani, K.A., Al-Tholaia, M.M.H., Nabus, H., Alih, S.C., Abdulkareem, Z., and Ganiyu, A. Comparative analysis of gradient-boosting ensembles for estimation of compressive strength of quaternary blend concrete. International Journal of Concrete Structures and Materials, 18, 18 (2024). https://doi.org/10.1186/s40069-023-00653-w

22 Hancock, J.T., and Khoshgoftaar, T.M. CatBoost for big data: an interdisciplinary review. Journal of Big Data, 7, 94 (2020). https://doi.org/10.1186/s40537-020-00369-8

23 Pemila, M., Pongiannan, R.K., Narayanamoorthi, R., AboRas, K.M., and Youssef, A. Application of an ensemble CatBoost model over complex dataset for vehicle classification. PLOS ONE, 19, e0304619 (2024). https://doi.org/10.1371/journal.pone.0304619

24 Sheridan, R.P., Wang, W.M., Liaw, A., Ma, J., and Gifford, E.M. Extreme gradient boosting as a method for quantitative structure–activity relationships. Journal of Chemical Information and Modeling, 56, 2353–2360 (2016). https://doi.org/10.1021/acs.jcim.6b00591

25 Shwartz-Ziv, R., and Armon, A. Tabular data: deep learning is not all you need. Information Fusion, 81, 84–90 (2022). https://doi.org/10.1016/j.inffus.2021.11.011

26 McElfresh, D. et al. When do neural nets outperform boosted trees on tabular data? arXiv:2305.02997 (2023). https://doi.org/10.48550/arXiv.2305.02997

27 Malashin, I.P., Tynchenko, V.S., Nelyub, V.A., Borodulin, A.S., and Gantimurov, A.P. Estimation and prediction of the polymers' physical characteristics using the machine learning models. Polymers, 16, 115 (2024). https://doi.org/10.3390/polym16010115

28 Kucukarslan, O. Machine learning-based heat deflection temperature prediction and effect analysis in polypropylene composites using CatBoost and Shapley Additive Explanations. Engineering Applications of Artificial Intelligence, 120, 106873 (2023). https://doi.org/10.1016/j.engappai.2023.106873

29 Phoeuk, M. et al. Accuracy prediction of compressive strength of concrete incorporating recycled aggregate using ensemble learning algorithms: multinational dataset. Advances in Civil Engineering, 2023, 5076429 (2023). https://doi.org/10.1155/2023/5076429

30 Uddin, M.N., Shanmugasundaram, N., and Praveenkumar, S. Prediction of compressive strength and tensile strain of engineered cementitious composite using machine learning. International Journal of Mechanics and Materials Design, 20, 671–716 (2024). https://doi.org/10.1007/s10999-023-09695-0

31 Khodadadi, N., Roghani, H., De Caso, F., El-kenawy, E.-S.M., Yesha, Y., and Nanni, A. Data-driven PSO-CatBoost machine learning model to predict the compressive strength of CFRP-confined circular concrete specimens. Thin-Walled Structures, 198, 111763 (2024). https://doi.org/10.1016/j.tws.2024.111763

32 Aydın, Y., Cakiroglu, C., Bekdaş, G., and Geem, Z.W. Explainable ensemble learning and multilayer perceptron modeling for compressive strength prediction of ultra-high-performance concrete. Biomimetics, 9, 544 (2024). https://doi.org/10.3390/biomimetics9090544

33 Lundberg, S.M., and Lee, S.-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4766–4777 (2017). https://doi.org/10.48550/arXiv.1705.07874

34 Tsioptsias, C. et al. Simulation and experimental study of the isothermal thermogravimetric analysis and the apparent alterations of the thermal stability of composite polymers. Polymers, 16, 1454 (2024). https://doi.org/10.3390/polym16111454

35 Gurnani, R., Kuenneth, C., Toland, A., and Ramprasad, R. Polymer informatics at scale with multitask graph neural networks. Chemistry of Materials, 35, 1560–1567 (2023). https://doi.org/10.1021/acs.chemmater.2c02991

36 Frie, C., Schork, N., Reiff, A., and Melz, T. Exploration of materials fatigue influence factors using interpretable machine learning. Fatigue and Fracture of Engineering Materials and Structures, 47, 2752–2773 (2024). https://doi.org/10.1111/ffe.14315

37 Elaskalany, M. et al. Stochastic multiscale modeling of electrical conductivity of carbon nanotube polymer nanocomposites: an interpretable machine learning approach. Advanced Engineering Materials, 26, 2401233 (2024). https://doi.org/10.1002/adem.202401233

38 Mykytyshyn, A. et al. Machine learning approaches for classification of composite materials. Modelling, 6, 118 (2025). https://doi.org/10.3390/modelling6040118

39 Wan, Z., Chen, S., Feng, X. & Sun, Z.-Y. From processing to properties: enhancing machine learning models with microstructural information in polymer nanocomposites. Composites Communications, 51, 102072 (2024). https://doi.org/10.1016/j.coco.2024.102072

40 Champa-Bujaico, E., Díez-Pascual, A.M., Redondo, A.L., and Garcia-Diaz, P. Interpretable machine learning framework to predict the glass transition temperature of polymers. Polymers, 16, 1049 (2024). https://doi.org/10.3390/polym16081049

41 Liu, B., Vu-Bac, N., Zhuang, X., Fu, X., and Rabczuk, T. Stochastic integrated machine learning based multiscale approach for the prediction of the thermal conductivity in carbon nanotube reinforced polymeric composites. Composites Science and Technology, 224, 109425 (2022). https://doi.org/10.1016/j.compscitech.2022.109425

[1*] Изтлеуова Б.,
докторант, ORCID ID: 0009-0009-7588-8976,
[*]e-mail: b.iztleuova@zhubanov.edu.kz
[1]Иманчиев А.,
ф.-м.ғ.к., қауымдастырылған профессор, ORCID ID: 0000-0002-1835-2501,
e-mail:imanchiev_ae@mail.ru
[1]Бекешев А.,
ф.-м.ғ.к., қауымдастырылған профессор, ORCID ID: 0000-0002-7038-4631,
e-mail: amirbek2401@gmail.com
[1]Жантурина Н.,
PhD, қауымдастырылған профессор, ORCID ID: 0000-0001-9540-6334,
e-mail: nzhanturina@mail.ru
[2]Усенкулова Ш.,
PhD, қауымдастырылған профессор, ORCID ID: 0000-0001-9417-755X,
e-mail: Sholpan1990@gmail.com
[3]Ванг С.,
PhD, қауымдастырылған профессор, ORCID ID: 0000-0001-5881-4400,
e-mail: wxcmx@ustc.edu.cn

[1] Қ. Жұбанов атындағы Ақтөбе өңірлік университеті,
Ақтөбе қ., Қазақстан
[2]К. Құлажанов атындағы Қазақ технология және бизнес университеті,
Астана қ., Қазақстан
[3]Қытай Ғылым және технология университетінің Өрт қауіпсіздігі
жөніндегі мемлекеттік негізгі зертханасы,
Хэфэй қ., Қытай

МИНЕРАЛ–ПОЛИМЕР КОМПОЗИТТЕРІНДЕГІ ТЕРМИЯЛЫҚ ЫДЫРАУДЫ БОЛЖАУ ҮШІН ГРАДИЕНТТІ БУСТИНГ ӘДІСТЕРІН ЭМПИРИКАЛЫҚ САЛЫСТЫРМАЛЫ ТАЛДАУ

Аңдатпа

Бұл зерттеуде минералдық толтырғыштары бар винил эфир матрицалары мен эпоксидтер негізіндегі композиттердің термогравиметриялық талдау (TGA) деректеріне қолданылған градиентті бустинг алгоритмдеріне математикалық және эмпирикалық талдау ұсынылған. Градиентті бустинг алгоритмі TGA-ның сандық мысалы негізінде кезең-кезеңімен көрсетілген. Одан кейін XGBoost және CatBoost алгоритмдері формалды түрде талданды. Рецензияланған әдебиеттерден алынған 478 дерек бойынша 5 рет қайталанған 5-қатпарлы кросс-валидация әдісімен 12 алгоритм салыстырылды. Extra Trees және CatBoost ең жоғары нәтижелерге статистикалық тұрғыдан тең деңгейде жетті (орташа $R^2 = 0.7430 \pm 0.0879$ және 0.7313 ± 0.0793 сәйкесінше; Nadeau & Bengio corrected $t = 0.483$, $p = 0.634$), және олар барлық сызықтық базалық модельдерден, SVR және MLP әдістерінен жоғары нәтиже көрсетті. Практикалық қолдану үшін CatBoost ұсынылады: оның реттелген бустингі шағын деректер жиындарында ($n < 1000$) бейтарап градиент бағалауын қамтамасыз етеді, ал градиент ағашының құрылымына негізделген SHAP талдауы TGA механизмдерімен үйлесімді физикалық тұрғыдан интерпретацияланатын белгі атрибуцияларын ашады. SHAP мәндері TGA-ның шың температурасы (T_max) мен матрица түрі негізгі анықтаушы факторлар екенін көрсетті. Ұсынылған алдын ала өңдеу құбыры, оның ішінде phr-дан wt%-ға түрлендіру, label encoding және медианалық толықтыру, әртекті TGA әдеби деректер жиындары үшін қайта жаңғыртылатын стратегияны қамтамасыз етеді.

Түйін сөздер: градиентті бустинг, XGBoost, CatBoost, полимерлік композиттер, машиналық оқыту, термогравиметриялық талдау, реттелген бустинг, математикалық негіздер, алгоритм таңдау

[1*] Изтлеуова Б.,
докторант, ORCID ID: 0009-0009-7588-8976,
*e-mail: b.iztleuova@zhubanov.edu.kz
[1]Иманчиев А.,
к.ф.-м.н, ассоциированный профессор, ORCID ID: 0000-0002-1835-2501,
e-mail:imanchiev_ae@mail.ru
[1]Бекешев А.,
к.ф.-м.н, ассоциированный профессор, ORCID ID: 0000-0002-7038-4631,
e-mail: amirbek2401@gmail.com
[1]Жантурина Н.,
PhD, ассоциированный профессор, ORCID ID: 0000-0001-9540-6334,
e-mail: nzhanturina@mail.ru
[2]Усенкулова Ш.,
PhD, ассоциированный профессор, ORCID ID: 0000-0001-9417-755X,
e-mail: Sholpan1990@gmail.com
[3]Ванг С.,
PhD, ассоциированный профессор, ORCID ID: 0000-0001-5881-4400,
e-mail: wxcmx@ustc.edu.cn

[1]Актюбинский региональный университет им. К. Жубанова,
г. Актобе, Казахстан
[2]Казахский университет технологии и бизнеса им. К. Кулажанова,
г. Астана, Казахстан
[3]Государственная ключевая лаборатория пожарной науки,
Университет науки и технологий Китая,
Хэфэй, Китай

ЭМПИРИЧЕСКОЕ СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ ГРАДИЕНТНОГО БУСТИНГА ДЛЯ ПРОГНОЗИРОВАНИЯ ТЕРМИЧЕСКОГО РАЗЛОЖЕНИЯ В ПОЛИМЕРНЫХ КОМПОЗИТАХ С МИНЕРАЛЬНЫМИ НАПОЛНИТЕЛЯМИ

Аннотация

В данном исследовании представлен математический и эмпирический анализ алгоритмов градиентного бустинга, применяемых к данным термогравиметрического анализа (ТГА) композитов на основе винилэфирных и эпоксидных смол с минеральными наполнителями. Алгоритм градиентного бустинга пошагово представлен на численном примере ТГА. Далее формально проанализированы алгоритмы XGBoost и CatBoost. Сравнение 12 алгоритмов выполнено на основе 478 данных из рецензируемой литературы с использованием пятикратно повторенной 5-блочной кросс-валидации. Методы Extra Trees и CatBoost показали статистически эквивалентные наилучшие результаты (среднее $R^2 = 0.7430 \pm 0.0879$ и 0.7313 ± 0.0793 соответственно; Nadeau & Bengio corrected $t = 0.483$, $p = 0.634$), превзойдя все линейные базовые модели, SVR и MLP. Для практического применения рекомендуется CatBoost: его упорядоченный бустинг обеспечивает несмещенные оценки градиентов на малых выборках ($n < 1000$), а SHAP-анализ структуры градиентных деревьев выявляет физически интерпретируемые атрибуции признаков, согласующиеся с известными механизмами ТГА. Значения SHAP показали, что пиковая температура TGA (T_max) и тип матрицы являются основными определяющими факторами. Предложенный пайплайн предварительной обработки, включающий преобразование phr в wt%, label encoding и медианную импутацию, обеспечивает воспроизводимую стратегию для разнородных литературных TGA-данных.

Ключевые слова: градиентный бустинг, XGBoost, CatBoost, полимерные композиты, машинное обучение, термогравиметрический анализ, упорядоченный бустинг, математические основы, выбор алгоритма.