Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning-Assisted Prediction of Brilliant Green Removal from Water Using CaTiO 3 Photocatalysis: Bayesian Optimization, SHAP Explainability, and Robustness Assessment

Md. Abdul Hamid Khan

24mschem114@student.cuet.ac.bd

Chittagong University of Engineering and Technology

Md. Mujahidul Islam

Chittagong University of Engineering and Technology

Shohanur Rahman

Islamic University Bangladesh

Research Article

Keywords: Brilliant Green degradation, CaTiO3 photocatalysis, Wastewater treatment, Machine learning, Explainable artificial intelligence (SHAP)

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11436329/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Machine Learning-Assisted Prediction of Brilliant Green Removal from Water Using $CaTiO_3$ Photocatalysis: Bayesian Optimization, SHAP Explainability, and Robustness Assessment.

Md. Abdul Hamid Khan[a,b*], Md. Mujahidul Islam[a] , Shohanur Rahman[b]

[a] *Department of Chemistry, Chittagong University of Engineering and Technology, Chattogram-4349, Bangladesh*

[b] *Department of Applied Chemistry and Chemical Engineering, Islamic University Bangladesh, Kushtia-7003, Bangladesh.*

Corresponding Author: 24mschem114@student.cuet.ac.bd

Abstract

Brilliant Green (BG)-contaminated wastewater threatens aquatic ecosystems by limiting light penetration, disrupting photosynthesis, and exposing organisms to potentially toxic compounds, raising broader environmental and human-health concerns. This study presents an explainable machine-learning framework for predicting BG removal during $CaTiO_3$-assisted photocatalytic water treatment. Using 63 published experimental observations for the 2:7 precursor composition, six regression models Decision Tree, Random Forest, Extra Trees, Gradient Boosting, XGBoost, and CatBoost were evaluated through Bayesian hyperparameter optimization and repeated cross-validation. Although Extra Trees achieved marginally higher holdout accuracy, CatBoost was selected for its superior repeated-validation performance, achieving a test R^2 of 0.9884, RMSE of 3.4629 percentage points, and mean cross-validation R^2 of 0.9697. SHAP analysis identified degradation time (60.57%), initial BG concentration (32.12%), and catalyst dosage (7.31%) as the principal contributors to model predictions. Feature-combination analysis reinforced this hierarchy, while CatBoost retained a test R^2 of 0.9778 under 10% Gaussian noise. Predicted BG removal (89.32%) closely matched the reported experimental value (88.46%). These findings provide a promising foundation for future experimental validation, broader catalyst screening, and interpretable optimization of photocatalytic wastewater treatment.

Keywords: Brilliant Green degradation; $CaTiO_3$ photocatalysis; Wastewater treatment; Machine learning; Explainable artificial intelligence (SHAP).

1. Introduction

Synthetic-dye pollution has become a persistent challenge for industrial wastewater treatment, particularly in regions where textile, leather, paper, printing, pharmaceutical, cosmetic, and related manufacturing activities are concentrated. More than 100,000 commercial dye formulations have been reported, and global synthetic-dye production is commonly estimated at several hundred thousand tonnes each year, reflecting the enormous volume of these compounds that move through industrial water streams [1],[2],[3].The difficulty is that not all this dye remains attached to the intended product. About 10-15% will be lost during application, depending on the dye class and processing conditions. and discharged into the wastewater, thus creating a direct connection between industrial production and discharge into wastewater. When effluent treatment is not sufficient, receiving waters [4],(Singhetal.,2024),(Aragaw,2024a). Among these pollutants, Brilliant Green (BG) is a cationic triarylmethane dye used in colouring, staining, microbiological, and antiseptic applications. Its intense colour and chemical persistence are useful in practice, but the same properties make it troublesome once it reaches an aquatic environment [7],[8],[9]. The effect of dye-containing wastewater is not limited to visible discoloration. Highly coloured effluents reduce light penetration through the water bodies, suppress photosynthetic activity in algae and aquatic plants, alter oxygen dynamics, and can disrupt microbial and ecological processes in receiving O_2 from waters[10],[11].These effects can extend through the wider aquatic ecosystem. Release of synthetic dyes has been associated with impaired growth and reproduction, oxidative stress, developmental toxicity, hepatotoxicity, and genotoxic effects in aquatic organisms, while persistent parent compounds and degradation products may contribute to exposure further up the food chain [12]. BG itself has been described as a poorly behaved and biologically harmful dye, with reported concerns including irritation, respiratory and gastrointestinal effects, cellular oxidative damage, and possible mutagenic or carcinogenic responses. For that reason, removing BG from wastewater is not simply about restoring water clarity. It is part of the broader need to reduce persistent organic loading and protect aquatic-water quality before contaminated industrial effluent reaches natural systems[13].

Several technologies have been explored for dye-bearing wastewater, including adsorption, coagulation-flocculation, membrane separation, biological treatment, ozonation, Fenton chemistry, and other advanced oxidation processes. These approaches are useful, but they also come with practical limitations. Some separation-based methods mainly move the pollutant from water into sludge or another phase, while biological treatment can be slow for structurally persistent dyes and may be sensitive to wastewater composition [14],[15]. The use of photocatalysis

of semiconductors is attractive as it attempts to photochemically convert the contaminant rather than simply relocate it. Under proper irradiation conditions, the electrons and holes produced by the photogeneration process are involved in surface reactions that result in the formation of highly reactive species that can attack Complex organic chromophores and gradually decompose them to simpler ones [16]. This light-driven route can operate under relatively mild conditions, may use UV, visible, or solar radiation depending on catalyst design, and can reduce the secondary solid-waste burden associated with purely transfer-based treatment methods [17], [18]. For BG specifically, photocatalysis is already supported by a substantial experimental base. TiO_2-based treatment has shown that BG can undergo sequential N-de-ethylation and chromophore breakdown, demonstrating that the dye molecule itself can be chemically transformed rather than merely removed from solution by adsorption [19]. High BG-removal efficiencies have also been reported using TiO_2 and modified TiO_2 systems, including Sr-doped and B-doped materials, showing that catalyst engineering can substantially improve degradation behaviour [20],[21],[22]. The same pattern has been observed with other visible-light-active systems. Cu-Co-Ti mixed oxides and Fe-modified TiO_2 composites have both been applied successfully to BG degradation, indicating that effective treatment is not restricted to one photocatalyst family or one irradiation regime [23],[24]. ZnO-based systems provide further evidence and, importantly, show that photocatalytic performance changes noticeably with catalyst dosage and pollutant loading. This is especially relevant for water-treatment optimization, as it confirms high removal efficiencies. It depends not only on the catalyst but also on how the process is conducted [25],[26]. More recent studies using B-doped TiO_2 and In_2O_3/g-C_3N_4 continue to support photocatalysis as an active treatment route for BG while reinforcing the same practical lesson: catalyst chemistry and operating conditions remain decisive [27],[28],[29].

Among oxide photocatalysts, $CaTiO_3$ offers an attractive perovskite platform for water treatment because its structural and electronic characteristics can support light-induced redox reactions, while its behaviour can still be tuned through composition and synthesis conditions. For BG specifically, Ernawati et al. synthesized $CaTiO_3$ from eggshell-derived $CaCO_3$ and TiO_2 using precursor ratios of 1:1, 1:3, 2:5, and 2:7, and generated time-dependent BG-removal data under different catalyst and pollutant conditions [30]. This is important because it demonstrates that $CaTiO_3$ can reduce BG, and this dataset has the added value of showing a dose-response relationship. It also shows the effect of the catalyst dosage on treatment performance, initial BG concentration, and that a reaction system can be studied as a function of degradation time

and catalyst composition. It's not a degradation test, but a multivariable water-treatment problem.

Photocatalytic water treatment is essentially a multivariable process. An increase in the amount of the catalyst can provide a higher number of active sites; however, too high a loading can also cause turbidity, scattering, or other problems. Clumping together: the higher the initial dye concentration, the higher the number of pollutants that will be introduced, and the results can affect light. The degradation time, which is the time required to achieve the same degradation, reflects the ability of photocatalytic reactions to continue [31]. Because these effects can interact nonlinearly, one-factor-at-a-time experimentation can describe individual trends while still missing combinations that govern the overall treatment response. As the number of operating variables grows, experimental optimization also becomes increasingly time-consuming and resource-intensive, which has encouraged the use of data-driven methods in photocatalytic water treatment [32].

Machine learning (ML) provides a practical way to learn nonlinear relationships directly from experimental observations without forcing BG removal to follow a single predefined empirical equation. In photocatalysis, ML is increasingly being used for degradation prediction, process optimization, catalyst screening, and identification of influential operating variables [32]. Tree-based ensembles and boosting algorithms are particularly useful for structured experimental datasets because they can accommodate nonlinearities and interactions among predictors. Random Forest, Gradient Boosting, XGBoost, CatBoost, and related methods have therefore received considerable attention in photocatalytic modelling [33]. At the same time, a high R^2 value alone does not guarantee a reliable water-treatment model. Small experimental datasets can still be vulnerable to favourable train-test splits, overfitting, unstable model rankings, and black-box predictions that say little about why treatment efficiency changes [34]. Explainability tools such as SHAP are useful in this setting because they connect predictive performance with the direction and strength of individual process-variable effects.

Despite the growing use of ML in photocatalytic water treatment, an explainable and validation-centered ML framework for $CaTiO_3$-mediated BG removal remains largely unexplored. Existing experimental work provides the necessary degradation data, and recent ML studies demonstrate the value of prediction and optimization, but relatively little attention has been given to combining dataset suitability, repeated validation, interpretability, feature

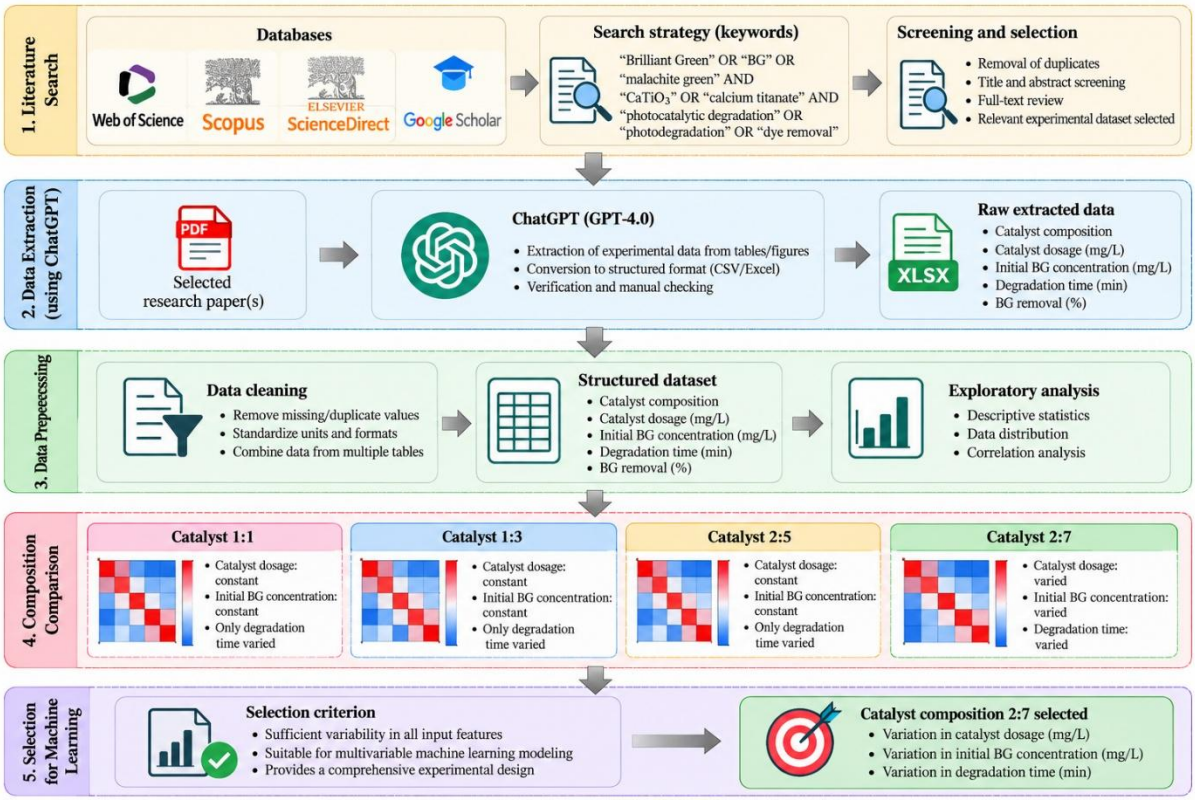
interaction, and noise robustness within this specific treatment system. The present study addresses that gap by reconstructing the published $CaTiO_3$-BG dataset and focusing multivariable modeling on the 2:7 composition, where catalyst dosage, initial BG concentration, and degradation time all vary meaningfully. Six tree-based regressors Decision Tree, Random Forest, Extra Trees, Gradient Boosting, XGBoost, and CatBoost are compared after Bayesian hyperparameter optimization, while repeated 5-fold cross-validation over 10 repetitions, holdout evaluation, overfitting-gap analysis, and model ranking are used to distinguish genuine generalization from a favourable single split. The framework then moves beyond routine accuracy comparison. Normalized feature importance and SHAP analysis are used to identify which operating variables actually drive the model response; all seven non-empty predictor combinations are evaluated to test whether that feature hierarchy is supported by predictive evidence; and Gaussian noise at 1, 2.5, 5, and 10% is introduced to examine whether the selected model remains stable when the input data are disturbed. Finally, the selected prediction is compared with the corresponding reported experimental BG-removal condition. The novelty of this work therefore lies not simply in applying ML to a photocatalytic dataset, but in integrating water-treatment prediction, repeated validation, interpretable process understanding, feature-level verification, and robustness testing within a single $CaTiO_3$ BG framework. The broader aim is to support a more reliable and data-efficient strategy for optimizing photocatalytic treatment of BG-contaminated water.

2. Materials and methods

2.1. Literature search, data source, and data extraction

The study was developed as a literature-data-driven machine-learning investigation rather than as a new experimental campaign. A structured search was conducted through Web of Science, Scopus, ScienceDirect, and Google Scholar to identify experimental studies reporting the photocatalytic degradation of Brilliant Green (BG) using $CaTiO_3$-based materials. Search combinations included the terms “Brilliant Green” OR “BG” OR “malachite green”, “$CaTiO_3$” OR “calcium titanate”, and “photocatalytic degradation” OR “photodegradation” OR “dye removal”. Duplicate records were removed, after which the titles and abstracts were screened, and potentially relevant articles were examined in full text. The experimental dataset reported by Ernawati et al. was ultimately selected because it provided time-dependent BG degradation data for four $CaCO_3/TiO_2$ precursor ratios (1:1, 1:3, 2:5, and 2:7) under systematically varied experimental conditions. The source study reported photocatalytic BG degradation for

irradiation times of up to 120 min [35]. ChatGPT (GPT-4.0) was used as an assisting tool during data extraction and organization. Numerical information reported in the selected publication was transcribed from tables and figures, organized into structured CSV/Excel-compatible records, and checked for consistency in variable names, units, and experimental conditions. ChatGPT was not used to generate missing experimental values or replace scientific judgment. The extracted records were manually cross-checked against the original publication before further statistical or machine-learning analysis. The complete literature-screening, AI-assisted extraction, preprocessing, and catalyst-selection workflow is summarized in Scheme 1.



Scheme 1. Schematic workflow of literature screening, AI-assisted data extraction, preprocessing, catalyst-composition comparison, and selection of the dataset for machine-learning analysis.

2.2. Data preprocessing and catalyst-composition selection

The extracted experimental data were organized into a structured dataset, with each row representing an individual BG-removal observation. Five variables were initially retained: catalyst composition, catalyst dosage (mg/L), initial BG concentration (mg/L), degradation time (min), and BG removal (%). Before model development, the dataset was examined for

missing or non-finite values, duplicate records, inconsistent units, and transcription errors. No synthetic observations were introduced during preprocessing.

The four $CaCO_3/TiO_2$ ratios of 1:1, 1:3, 2:5, and 2:7 were then investigated. Their appropriateness for multivariable modeling. In the 1:1, 1:3, and 2:5 subsets, catalyst dosage and initial BG concentration were not significantly altered, while degradation time was. Most of the experimental variation is accounted for by this factor. These subsets provided little information. For estimation of the individual effects of the three operating variables. In contrast, the 2:7 The variable conditions used in composition were catalyst dosage, initial BG concentration, and degradation. The proposed regression analysis is best suited to this subset of data, because it is the most representative for the time. The final dataset comprised 63 observations, covering catalyst dosages of 50-200 mg/L and initial BG. The degradation times ranged from 0 to 120 min, while the BG removal (%) ranged from 10 to 40. The prediction target was a continuous variable.

Distribution and variability were characterized through descriptive statistics. Experimental variables. Deviations from normality were evaluated using a Shapiro-Wilk test, and Spearman's rank correlation was used to examine monotonic associations between the Operating variables and BG removal. Since the available observations were not a complete factorial experimental design, treatment differences were checked separately within each matched experimental condition, comparable by Friedman test. Kendall's coefficient of to quantify the magnitude of these differences, concordance (W) was reported. Paired Wilcoxon signed-rank tests were conducted for post hoc comparisons as appropriate. Multiple testing was controlled by Holm adjustment.

2.3. Machine-learning modeling

Six tree-based regression algorithms were compared under the same modeling framework: Decision Tree (DT), Random Forest (RF), Extra Trees (ET), Gradient Boosting (GB), XGBoost (XGB), and CatBoost (CB). Since tree-based models were chosen because they can be trained on many features and are robust, they were the preferred models. have nonlinear response and interactions among experimental variables without needing to; A pre-established mathematical equation that relates the predictors to the removal of BG. XGBoost CatBoost use ordered categories, while scalable gradient-boosted decision trees use categories. Boosting: increases to achieve better generalization and a decrease in prediction shift in the case of sequential boosting [36],[37]. The 2:7 dataset was randomly divided into 80% training and 20% testing subsets

using random_state = 42. The training data were used for model construction and hyperparameter optimization, whereas the test data were kept separate for final holdout evaluation. Since all selected algorithms were tree-based, feature scaling was not applied to the three numerical predictors.

2.4. Hyperparameter optimization and validation

The hyperparameters of all six algorithms were optimized using Bayesian optimization with BayesSearchCV. For each model, 50 search iterations were performed using 5-fold cross-validation on the training data, with RMSE used as the optimization criterion. The search spaces included model-specific parameters such as tree depth, number of estimators, learning rate, minimum leaf size, sampling fraction, and regularization terms. The best hyperparameter combination identified for each algorithm was retained for final fitting. To reduce the dependence of model selection on a single favourable data partition, the tuned models were subsequently evaluated using 5-fold cross-validation repeated 10 times. The repeated-validation results were considered together with the independent test-set results during final model comparison.

2.5. Performance assessment

Model performance was assessed using six complementary regression metrics: the coefficient of determination (R^2), mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), median absolute error (MedAE), and mean absolute percentage error (MAPE).

The coefficient of determination was calculated as:

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

The mean squared error was expressed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (2)$$

and RMSE as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (3)$$

The mean absolute error was calculated using:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - \hat{y}_i \,| \qquad (4)$$

while the median absolute error was:

$$\mathrm{MedAE} = \mathrm{median}(|\, y_i - \hat{y}_i \,|) \qquad (5)$$

MAPE was expressed as:

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=0}^{n} \frac{(y_i - \hat{y}_i)}{y_i} \qquad (6)$$

where y_i, $\hat{y}_i$, and n denote the experimental value, predicted value, mean experimental value, and number of observations, respectively. Because the dataset included zero BG-removal observations, the computational implementation used in the analysis was retained for handling zero denominators rather than applying Eq. (6) directly where $y_i = 0$. Models were not ranked from R^2 alone. Actual-versus-predicted plots, residual behaviour, train-test differences, repeated cross-validation, and overall metric ranking were considered together to assess predictive stability and possible overfitting.

2.6. Model interpretation, feature analysis, and robustness assessment

Normalized feature importance was first obtained for all six models to compare the relative influence of catalyst dosage, initial BG concentration, and degradation time. The selected best-performing model was subsequently interpreted using SHAP (SHapley Additive exPlanations), which assigns a contribution value to each predictor for individual predictions and therefore provides information on both feature importance and the direction of its influence [38]. To test whether the SHAP-derived importance pattern was supported by predictive evidence, CatBoost was additionally evaluated using all seven non-empty combinations of the three input variables. Model robustness was then examined by introducing controlled Gaussian perturbations at 1, 2.5, 5, and 10% noise levels and tracking the corresponding changes in R^2, MSE, RMSE, MAE, MedAE, and MAPE. Finally, the selected operating condition was restricted to the experimental domain represented by the published dataset, and the CatBoost prediction was compared directly with the corresponding reported experimental BG removal rather than extrapolated beyond the available data range.

3. Result and discussion

3.1. Descriptive statistics

Before developing the machine-learning models, the $CaTiO_3$ (2:7) dataset was examined to understand the distribution of experimental conditions and the variation in Brilliant Green (BG) removal efficiency. The dataset comprised 63 observations, covering catalyst dosages of 50-200 mg/L, initial BG concentrations of 10-40 mg/L, and degradation times of 0-120 min. BG removal ranged from 0 to 89.53%, with a mean efficiency of 51.55%, indicating considerable variation in the treatment response across the investigated conditions. This broad response range is particularly relevant for predictive modeling, as it allows the models to learn from both low- and high-removal observations rather than from a narrow range of treatment efficiencies. The variability also reflects the importance of considering operating conditions when evaluating photocatalytic wastewater treatment. Catalyst dosage, initial pollutant concentration, and irradiation time are widely recognized as important factors influencing photocatalytic degradation, although their relative contributions depend on the catalyst system and experimental conditions. Examining their distributions and relationships with BG removal therefore provides a useful foundation for interpreting the subsequent machine-learning results.

Table 1.

Statistical summary of the variables in the $CaTiO_3$ (2:7) dataset.

Statistic	Catalyst dosage (mg/L)	Initial BG concentration (mg/L)	Degradation time (min)
Count	63	63	63
Mean	92.857	18.571	60.000
Standard deviation	56.695	11.339	39.041
Variance	3214.286	128.571	1524.194
Minimum	50.000	10.000	0.000
25th percentile	50.000	10.000	30.000
Median (50%)	50.000	10.000	60.000
75th percentile	150.000	30.000	90.000
Maximum	200.000	40.000	120.000
Range	150.000	30.000	120.000
IQR	100.000	20.000	60.000
CV (%)	61.056	61.056	65.068
Skewness	0.907	0.907	0.000
Kurtosis	-0.737	-0.737	-1.232
Correlation with target	0.342	-0.467	0.818

The three predictors showed significant variation across the (Table 1) selected $CaTiO_3$ (2:7) dataset. Degradation time showed the highest relative variability, with the coefficient of variation (CV) for catalyst dosage and initial BG concentration being 10.03% and 12.30%, respectively. each had a CV of 61.06%. The BG-removal response was also broadly distributed

(CV = These represent significant variations between treatment efficiencies (56.14%). investigated conditions. The distributional characteristics gave further insight into the dataset. The degradation time was nearly symmetric (skewness = 0), whereas the catalyst dosage and initial BG concentration had positive skewness (0.907), reflecting a high concentration of the data in the upper tail. Clumping of observations at lower values. BG removal demonstrated a slight improvement. The skewness was negative (-0.426), indicating a slightly longer tail towards lower removal efficiencies. Excess kurtosis was negative for all four variables, suggesting relatively flat distributions. Following a normal distribution. All these traits indicate that the data set measures Significant fluctuation in both operating conditions and photocatalytic treatment response. Such variation is a good thing for the study of nonlinear regression because it can take care of changes in the removal of BG in various experimental conditions. Despite the uneven distribution of model performance, it should be taken with some grain of caution, especially when used in less familiar settings. High-resolution data samples in the experimental domain [39]. The Pearson correlation heatmaps. Figure 1 also illustrates further relationships between the experimental variables and BG removal efficiency. Degradation time was found to be as follows for the $CaTiO_3$ (2:7) composition. A longer irradiation time correlated best with BG removal (r = 0.818), meaning the longer the irradiation time, The treatment efficiencies for the periods were generally high in the investigated experimental range. A weaker positive correlation (r = 0.342) was found with the dosage of catalyst; this implied that the more photocatalyst added, the better the BG removal. Removal, although this relationship was less pronounced than that observed for degradation time. Initial BG concentration, however, showed a moderate negative correlation with removal. efficiency (r = -0.467), indicating that higher dye concentrations were generally associated with a lower percentage of BG removed. This is typical behaviour of the kind one would expect to see. Observed in heterogeneous photocatalytic treatment, increasing pollutant concentration. can compete with available surface-active sites and enhance dye absorption. molecules, which may decrease the number of photons hitting the catalyst surface. Under comparable conditions, these effects can affect the percentage of BG that can be degraded in a specified time. irradiation period. However, these reported correlations describe associations within the It is determined by available data and does not reflect any independent effect of the operating. variables [40]. The composition-wise correlation analysis also enabled the explanation of differences. It is noted that experimental coverage for the four $CaCO_3/TiO_2$ precursor ratios reported in the original $CaTiO_3$-BG study. For the 1:1, 1:3, and 2:5 compositions, most of the variation observed was due to the degradation time and the initial BG concentration, which

were found to be associated with degradation, and the dosage of the catalyst was not. largely unchanged. These subsets were helpful for the purposes of examining time-dependent BG removal; they were found to be less suitable for the other operating variables because they vary only slightly. Building a regression model with three predictors. The 2:7 composition, however, was used. The observations were recorded under various conditions of catalyst dosage, initial BG concentration and degradation time, which gave a more informative experimental basis for investigating their combined effects. Photocatalytic treatment efficiency and relationships. The 2:7 data set was therefore chosen based more on predictor variability and experimental coverage than on the strength of the system. Its correlation coefficients or the assumption of high photocatalytic power. This distinction is important because correlation is the association of variables but does not imply causation. It cannot mitigate lack of variation in the experimental data. The selection therefore retains the original $CaTiO_3$-BG experiment scope but also offers a more suitable basis for further multivariable machine learning analysis [41].

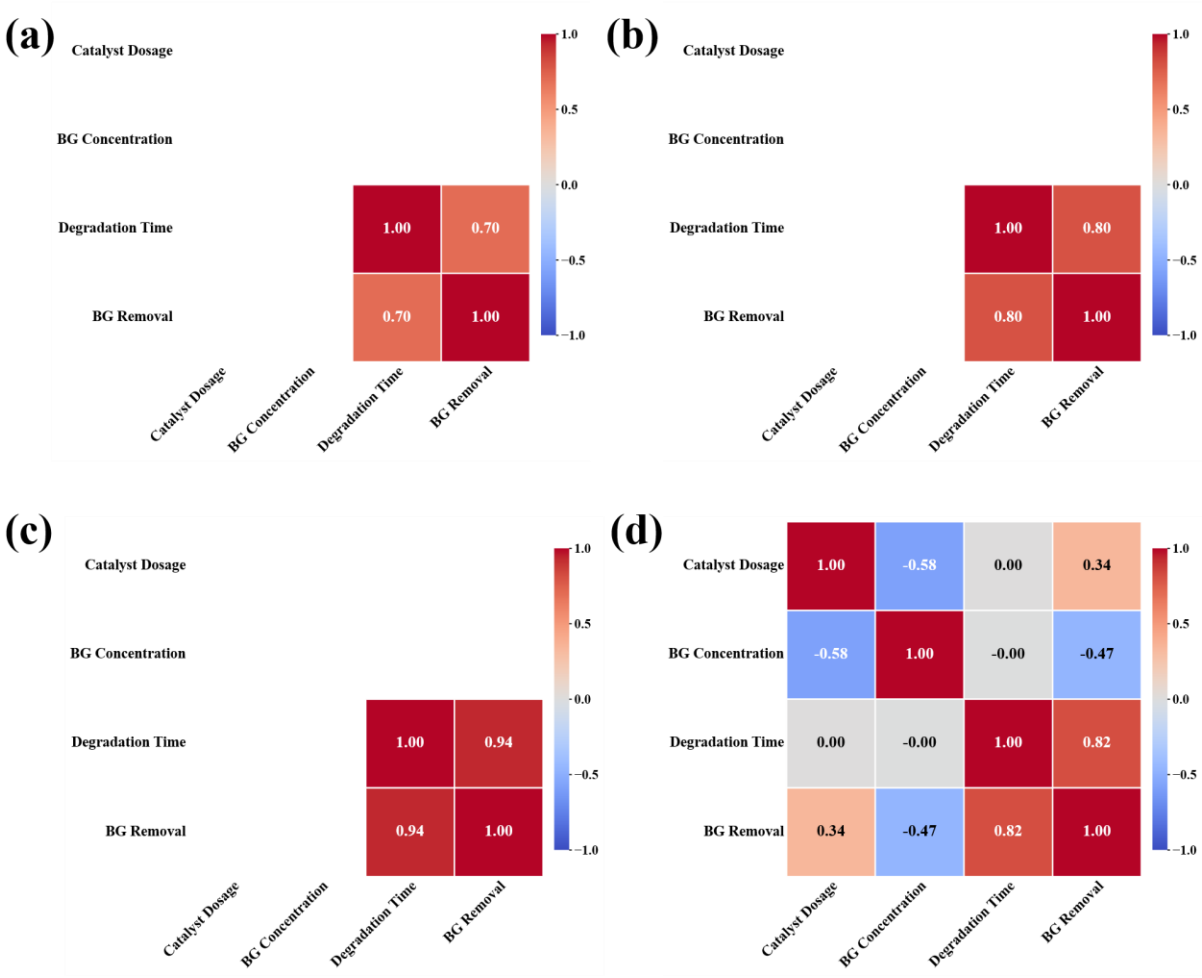


Fig. 1. Pearson correlation heatmaps showing the relationships among experimental variables and BG removal for the $CaTiO_3$ catalyst compositions of 1:1, 1:3, 2:5, and 2:7.

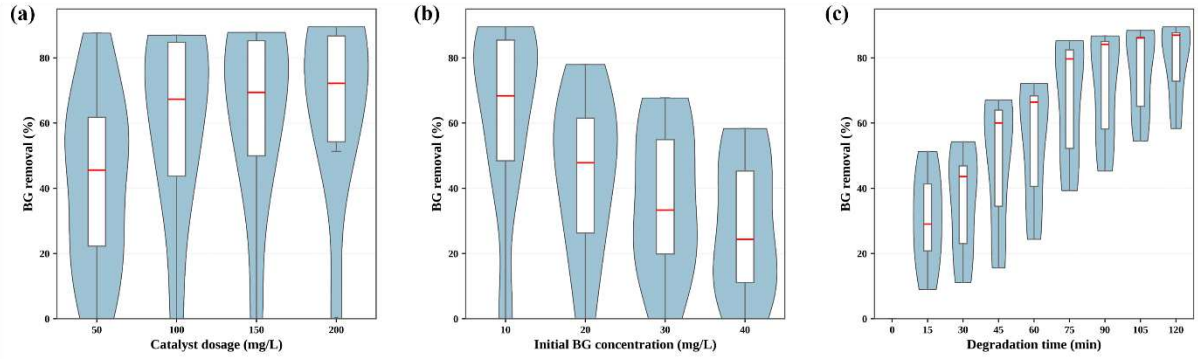


Fig. 2. Violin-boxplots showing the distribution of BG removal across catalyst dosage, initial BG concentration, and degradation time for the selected $CaTiO_3$ (2:7) dataset.

The violin-boxplots in Fig. 2 also show these relationships. BG removal the degradation time generally had an increasing effect, and this trend was consistent with the strong positive correlation. This association was noted in the correlation analysis. By contrast, the initial BG concentration indicated the opposite trend, where higher concentrations were related to lower removal efficiencies. However, the treatment response was also influenced by catalyst dosage, as its effect on degradation time was not as strongly expressed. This comparatively weaker correlation coefficient is comparable to that of catalyst dosage. In terms of mechanism, a higher photocatalyst loading will result in a higher available surface area and number of active sites, and hence the improvement in the resulting process. Degradation is not necessarily linear, depending on the Optimum optimization of the catalyst loading, light penetration, and pollutant concentration. Overall, this is a major factor in shaping distribution patterns in Fig. 2. The primary sources of BG removal in the chosen $CaTiO_3$ (2:7) were concentration and dosage. The dosage of the catalyst had a secondary but not insignificant effect on the dataset's photocatalytic treatment performance [42]. Taken together, the descriptive statistics, Pearson the pattern across the different $CaTiO_3$ (2:7) dataset. The degradation time was highly positively correlated with BG removal; Initial BG concentration was inversely related to treatment efficiency. Catalyst Dosage showed a positive correlation, but it was less strong. Weaker in the range of experiments investigated. These findings highlight the importance of When assessing photocatalytic treatment, all three operating variables were taken into consideration. Instead of depending on just one factor, it is a performance. Notably, the 2:7 mixture yielded There is enough variation in catalyst dosages, initial BG concentrations, and degradation time to Unlike the other precursor compositions tested, it allows for multivariable regression. This experimental coverage, together with the observed variation in BG removal, provided the basis for developing and comparing the machine-learning models presented in the following section.

3.2. Model performance evaluation

To have a fair comparison of the six tree-based regression models, they were evaluated after tuning their model parameters using the Bayesian hyperparameter optimization method. The optimization process was not based on default parameters but rather aimed to find combinations of parameters to ensure a more balanced trade-off between prediction and generalization. This is especially valuable for the use of Bayesian Optimization, where successive evaluations of the hyperparameters build on previous evaluations and the search becomes increasingly focused on the most likely region of the hyperparameter space, rather than exploring all possible combinations. In recent years, similar optimization schemes have been proven to be effective in machine-learning studies of photocatalytic processes, especially in comparative studies of different ensemble algorithms in an identical predictive model [43]. The correlation between the experimental and predicted BG removal is plotted in Fig. 3. The best ensemble models yielded predictions that were in good agreement with the 1:1 reference line in both the training and testing sets. The agreement was closest for the CatBoost and Extra Trees models, and the agreement was good overall for the Gradient Boosting and XGBoost models throughout most of the removal range observed. There was a slight increase in the spread of the data for Random Forest, whereas the decision tree had the most obvious deviation from the ideal line. As is common with ensemble learning, by combining multiple trees, this pattern will typically yield more flexibility in capturing the non-linear relationship than a single decision structure. If you are working with photocatalytic modelling, the reaction time, pollutant concentration, and catalyst quantity may be the parameters you are interested in. This additional flexibility can be very useful if dosage is related in a non-linear fashion. Likewise, ensemble and boosting methods have been shown to provide improved performance over simpler tree-based approaches. been reported via photocatalytic dye degradation machine learning research [44].

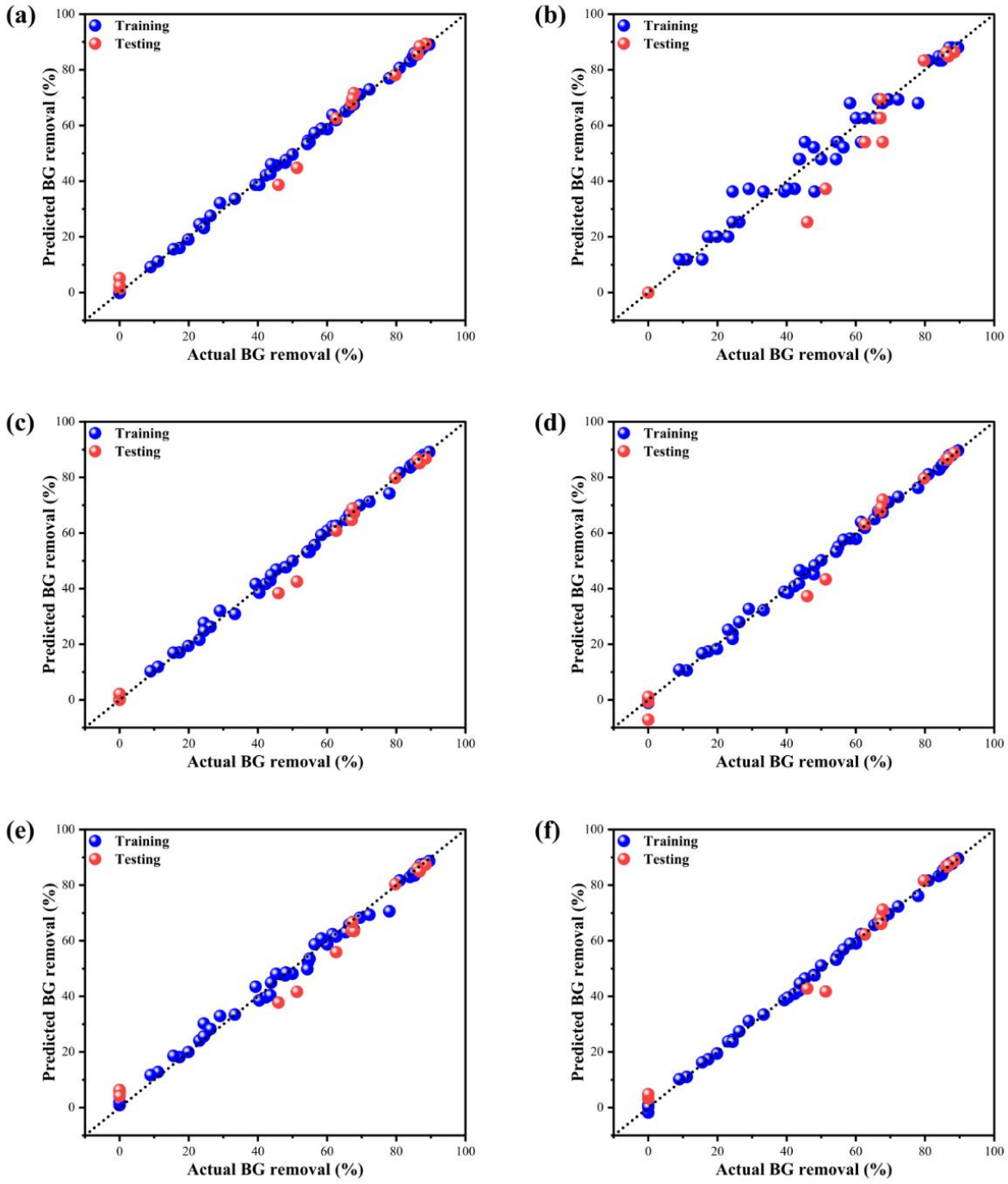


Fig. 3. The scatter plots show the training and testing predictions of BG removal (%): (a) CatBoost, (b) Decision Tree, (c) Extra Trees, (d) Gradient Boosting, (e) Random Forest, and (f) XGBoost.

The residual plots, provided in the Supplementary, offer a second view of model behaviour by showing how far individual predictions deviated from their corresponding experimental values. The better-performing models produced comparatively compact residual distributions around

zero, whereas the weaker models showed broader deviations. This is helpful because models with identical R^2 values can have very different prediction errors in size and pattern. Residual inspection is thus a complement to traditional accuracy measures in that it identifies if the errors are reasonably balanced through the prediction range or if there is a systematic under- or overprediction. An equivalent set of predicted-versus-observed and residual plots has been employed in recent environmental and photocatalytic ML research to evaluate the reliability of the models, in addition to a single goodness-of-fit statistic [45].

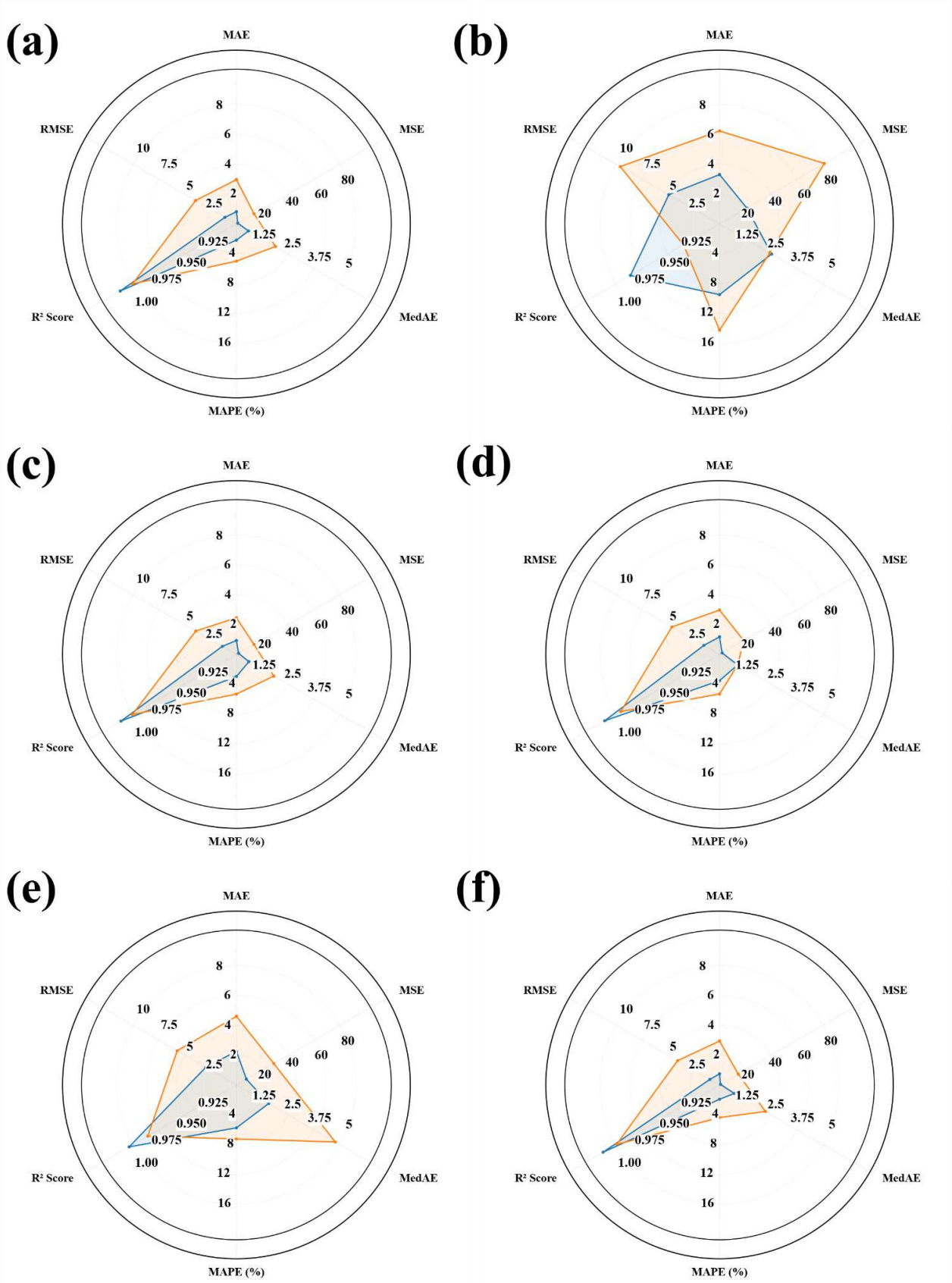
Table 2.

Comparative training and testing metric results of the six Bayesian-optimized machine-learning models.

Model	Dataset	R^2	MedAE	MAE	MSE	RMSE	MAPE
CatBoost	Training	0.998762	0.514244	0.722262	0.952217	0.975816	1.904893
CatBoost	Testing	0.988352	1.661787	2.615365	11.991408	3.462861	4.382315
Decision Tree	Training	0.975782	2.216880	2.909544	18.622693	4.315402	8.327778
Decision Tree	Testing	0.930707	2.136752	5.481591	71.337165	8.446133	12.517450
Extra Trees	Training	0.998139	0.529915	0.822607	1.430856	1.196184	2.598930
Extra Trees	Testing	0.988420	1.589744	2.157462	11.921712	3.452783	4.676320
Gradient Boosting	Training	0.997681	0.769637	1.035982	1.783269	1.335391	3.086001
Gradient Boosting	Testing	0.984041	0.771662	2.615999	16.430291	4.053430	4.660512
Random Forest	Training	0.991185	1.370182	1.979024	6.778332	2.603523	5.052540
Random Forest	Testing	0.975375	4.201727	4.042640	25.351102	5.034988	6.366183
XGBoost	Training	0.999129	0.626457	0.652625	0.669541	0.818255	1.690479
XGBoost	Testing	0.987597	1.951794	2.588225	12.769278	3.573413	3.842018

The numerical results obtained in Table 2 reveal that the top models had only minor differences. No single algorithm achieved the best results for all the error measures, and the results for margins were quite similar to those for the other error measures. Extra Trees achieved the highest test R^2 (0.988420) and the lowest test RMSE (3.452783). CatBoost followed almost identically, with a test R^2 of 0.988352 and an RMSE of 3.462861. XGBoost recorded the lowest test MAE (2.588225) and MAPE (3.842018%), whereas Gradient Boosting produced the smallest test MedAE (0.771662). Random Forest remained reasonably predictive but showed noticeably larger errors, while the Decision Tree was the weakest overall, with a test R^2 of 0.930707 and an RMSE of 8.446133. These differences illustrate why model performance is

414 better judged using several complementary metrics: R^2 describes explained variance, while
415 MAE, RMSE, MedAE, and MAPE capture different dimensions of prediction error [46]. The
416 training and testing profiles are summarised further in the radar plots displayed in Fig. 4.
417 CatBoost, Extra Trees, Gradient Boosting, and XGBoost showed relatively even performance.
418 The differences between profiles were more evident for the larger train-test difference in the 6
419 metrics. Weaker models. These contrasts are easier to interpret in the radar representation since
420 Goodness of fit and multiple error measures can be taken as a whole and not as separate items.
421 isolation. Importantly, the holdout results alone do not warrant declaring the final best model
422 at this stage. Extra Trees outperforms CatBoost by a slight margin for test R^2 and RMSE, being
423 almost tied in the third one. It is not possible to distinguish on the same holdout set. The general
424 question of which model generalizes most consistently is thus investigated later by repeated
425 cross-validation and the overfitting gap. Overall model ranking analysis.

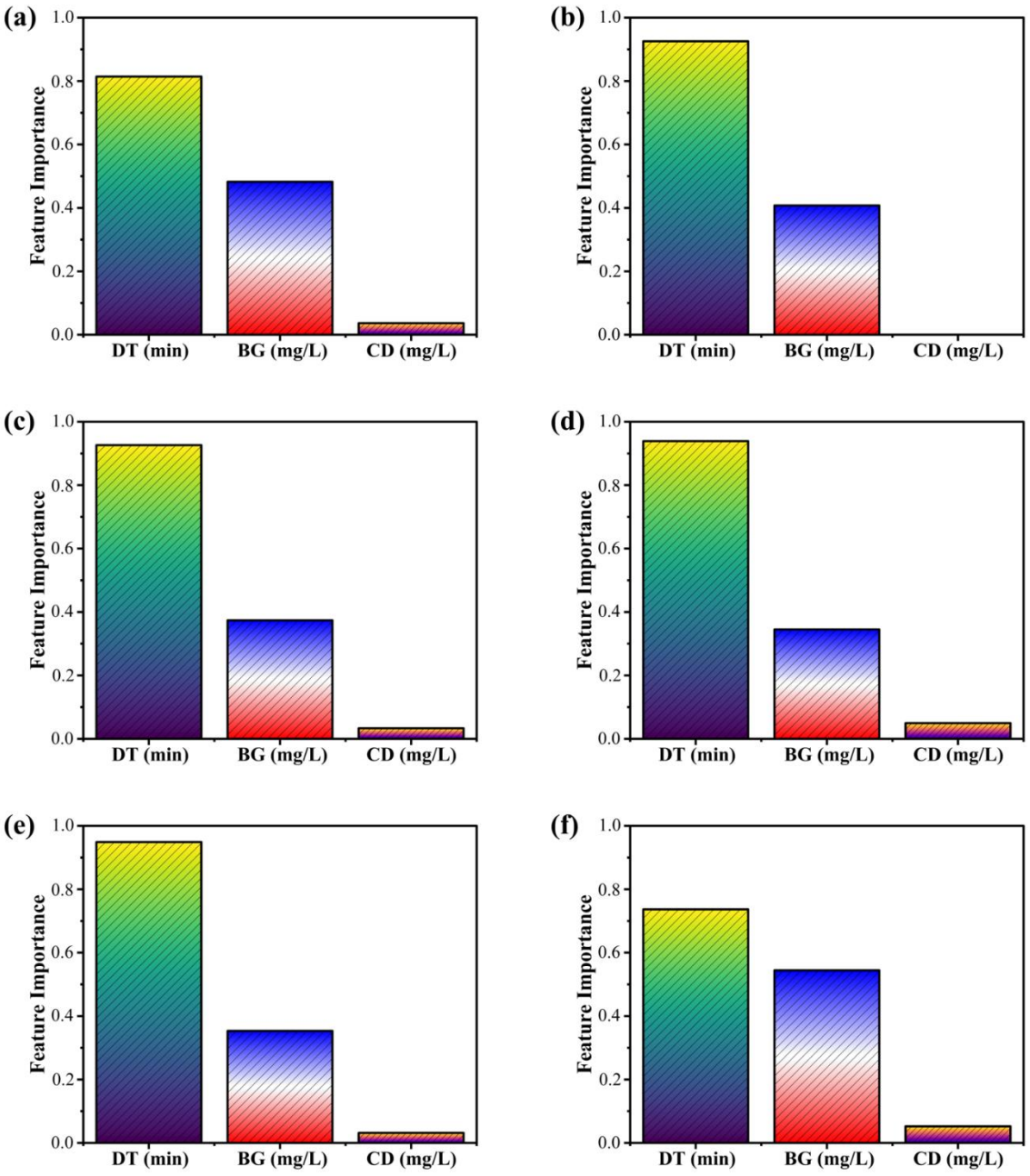


426

Fig. 4. Radar plots comparing the training and testing performance of the six machine-learning models using R^2, MedAE, MAE, MSE, RMSE, and MAPE.

3.3. Normalized importance of factors under the models

The relative contribution of the three experimental variables was tested for all six. To compare the algorithms for the removal of BG, regression models were used to assess whether they were using the same set of predictors. These normalized feature-importance profiles are shown in Fig. The values given to each of the variables varied between the different models, but A constant trend was seen degradation time was the most prominent factor for most. The order of variables that were found to affect the rate were: models, followed by initial BG concentration, and then followed by catalyst dosage generally. Smallest contribution. This overall ranking is in line with the statistical analysis in Section 3.1; degradation time had the highest correlation with BG removal. This agreement strengthens the correlation analysis with the model-derived importance profiles. Interpretation: irradiation duration was the main factor influencing the response in the selected 2:7 experimental domain.



441

442 **Fig. 5.** Normalized feature importance of catalyst dosage, initial BG concentration, and
443 degradation time across the six machine-learning models: (a) CatBoost, (b) Decision Tree, (c)
444 Extra Trees, (d) Gradient Boosting, (e) Random Forest, and (f) XGBoost.

445 Photocatalytic removal is a process that, if continued, generates a new set of energy levels in
446 the object. Photogenerated charge carriers and other species reduce the dye, thereby increasing
447 the likelihood of reduction. The molecules to be broken down. The effect of time also has been
448 quite evident in recent times; data-driven photocatalytic studies. In an experimentally proven

ML investigation of cationic Among the most important factors that were identified were dye degradation, reaction time and the initial dye concentration. In other words, the sensitive process variables were identified. In a separate study of green-synthesized photocatalysts, it was found that many of them exhibited that the longest irradiation time was the most significant positive parameter controlling the degradation performance [47]. Initial BG concentration was always perceived as to be of medium priority in the importance profiles. This is what you would expect, since the amount of dye that will stick is directly related to the amount of starting pollutant. must be degraded under a fixed set of reaction conditions and can change both surface competition and light availability in the suspension. Other feature prioritization patterns have been observed to be like the one observed here. This has been reported in general water-treatment datasets of photocatalytic studies and is shown in the following chart as a function of pollutant concentration. The catalyst-related variables were among the most influential ones, together with others [48]. The dosage of the catalyst had less of an effect in this dataset than the other two predictors and therefore was not as strong, while its importance was still not negligible across the models. This weaker contribution should not be taken as an indication that the loading of the catalyst is unimportant in photocatalysis; Instead, it is an indication of the relative impact of it within the range and structure of the 2:7 data set. The importance given to a variable always depends on the space of the experiment. The model other photocatalytic studies carried out at various ranges of catalyst and dye have, for instance, determined that the most important parameters for the degradation efficiency are catalyst dosage and dye concentration [49]. Overall, the feature-importance analysis indicates that the six models learned a somewhat similar hierarchical structure of experimental influence, although they differed in terms of the internal tree structure. The degradation time continued to be the most important factor, with initial BG concentration contributing significantly to second place and catalyst dosage in third, but not insignificant. The values of the importances of the different models are given in the Supplementary, and the general trend of these importances will be explored further in the following sections using the selected best model and SHAP analysis.

3.4. Comparative performance evaluation and model selection

To differentiate the accuracy of training from real predictive stability, a direct comparison of the six models was conducted. This behaviour is most easily seen in the training-testing R^2 gap in Fig. 6(a). The biggest difference was found in Decision Tree, with about 0.045 difference between training fit and test fit. Random Forest also had a slightly larger but significant difference of ~0.016. Extra Trees, CatBoost, XGBoost, and Gradient Boosting, on the other

hand, had much closer train-test gaps of around 0.010, 0.010, 0.012, and 0.014. The smaller the gap is, the better a model will fit the data it was trained on and the less it will deviate from the actual data that it is applied to. But remember the gap should be read alongside the absolute predictive accuracy, not by itself. Fig. 6 shows the second panel that displays the average model ranking together with the test MAE, thus providing a combined view of accuracy and overall consistency. This comparison is especially relevant for this instance where the top models were very similar with respect to holdout results in Table 2. The individual test R^2 and RMSE were best for Extra Trees, the lowest test MAE and MAPE for XGBoost, and the lowest test MedAE for Gradient Boosting. CatBoost, however, demonstrated the best overall repeated-validation behaviour, with the maximum mean repeated-CV R^2 of 0.969673, and the minimum mean CV RMSE of 3.904361. That is, it wasn't because of the best performance on all the holdout metrics, but because its performance was consistently strong on multiple partitions of the data. When there is only a small difference in results between two or more models on one train-test split, the distinction is significant because repeated cross-validation is a better way of assessing the stability of the difference. When it comes to the choice of ensemble algorithms among which cross-validated performances are similar, the same emphasis has been observed in recent photocatalytic ML studies [50]. The joint ranking thus led to CatBoost being selected as the most well-rounded model for further evaluation and improvement. interpretation. The test R^2 of 0.988352 was almost identical to the best holdout, and it wasn't far from the results of the holdout. The repeated-validation performance of the result obtained by Extra Trees is 0.988420. stronger. CatBoost also kept a small train-test R^2 gap, which is around 0.0104, which is a strong argument for the CatBoost model. That it had good predictive power, which was not limited to the training observations. The model’s ability to keep up performance from one partition to another is especially important for a small number of observations, and when only a single split can be put into one of the two sets, use the present dataset. Excessive weight is given to a few test samples. Recent catalyst-screening studies have also observed high generalization from CatBoost when nonlinear relationships are observed. All the material and operating variables must be modeled at the same time.

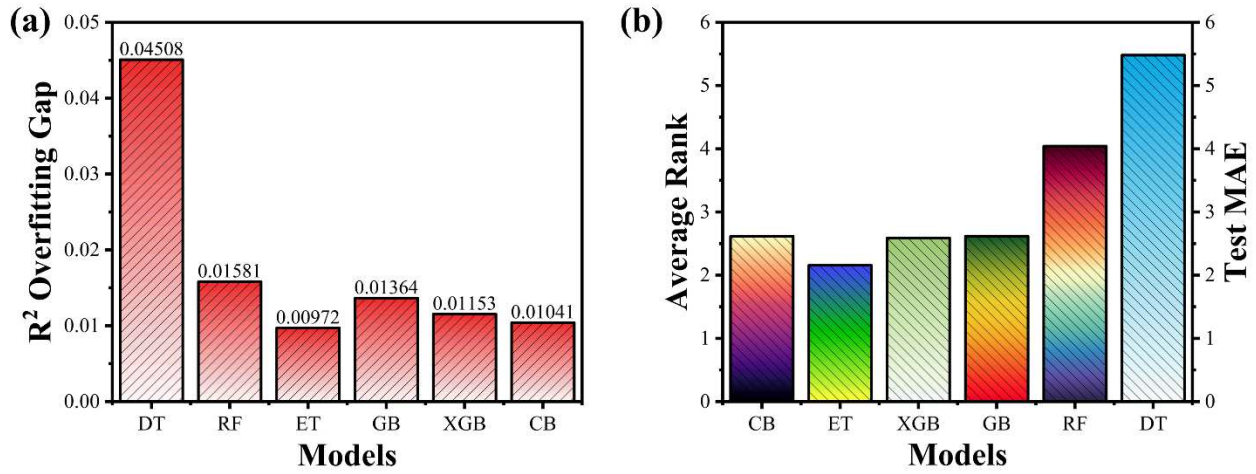


Fig. 6. Comparative performance of the six machine-learning models: (a) R^2 overfitting gap between the training and testing datasets and (b) average model ranking with the corresponding test MAE values.

Taken together, Fig. 6 shows why the final model was not selected solely from the highest holdout R^2. Extra Trees had a marginal numerical advantage on that single test partition, but CatBoost provided a stronger balance between holdout accuracy, repeated-validation performance, and train-test consistency. It was therefore chosen as the final model for the subsequent interpretation and robustness analysis, and prediction under best scenario analysis. experimental condition.

3.5. SHAP-based feature analysis of the best model

CatBoost was selected for further interpretation because it demonstrated the most consistent overall performance across the model evaluations. SHAP analysis was subsequently employed to examine the relative contributions of the three operating variables to predicted Brilliant Green (BG) removal. As illustrated in Fig. 7(a), degradation time was the most influential predictor, accounting for 60.57% of the total global SHAP importance, followed by initial BG concentration (32.12%) and catalyst dosage (7.31%). The beeswarm plot in Fig. 7(b) provided further insight into the direction of these relationships. Longer degradation times were generally associated with positive SHAP values, indicating higher predicted BG removal, whereas increasing initial BG concentration tended to shift predictions toward lower removal efficiencies. Catalyst dosage showed a comparatively smaller range of SHAP values, consistent with its lower contribution to the model predictions. Overall, these findings suggest that degradation time and initial pollutant concentration played the largest roles in shaping the

CatBoost predictions for $CaTiO_3$-assisted BG wastewater treatment within the investigated experimental range [51].

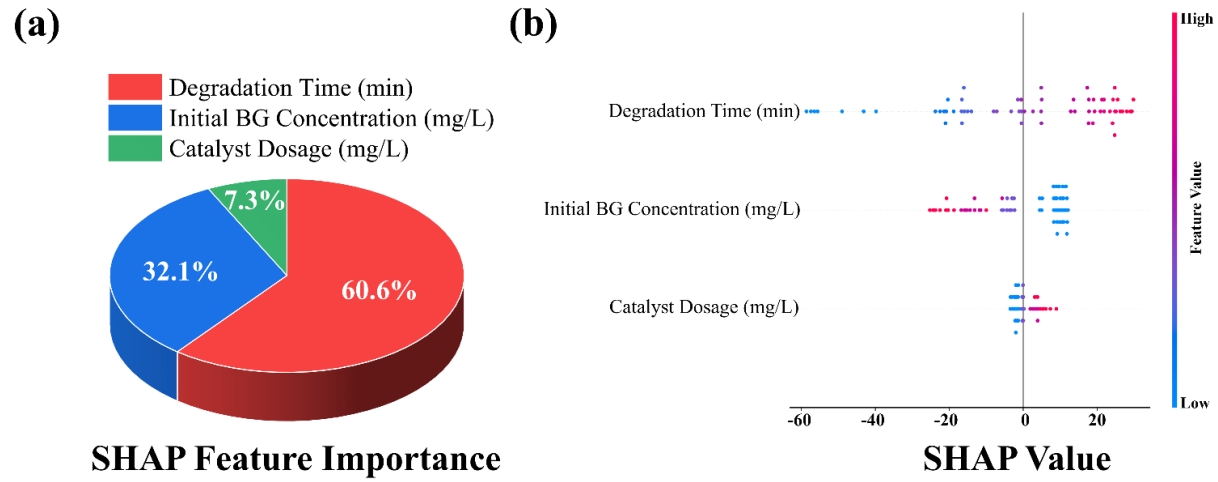


Fig. 7. SHAP analysis of the CatBoost model: (a) global percentage contribution of the input variables and (b) beeswarm plot illustrating the magnitude and direction of their effects on predicted BG removal.

To investigate the SHAP-derived feature ranking, CatBoost was tested with all seven All non-empty combinations of the three input variables (Table 3). Degradation time alone achieved a test R^2 of 0.8287, whereas catalyst dosage and initial BG concentration individually showed much weaker predictive performance. Combining degradation time with catalyst dosage increased the test R^2 to 0.9005, while pairing it with initial BG concentration improved performance more substantially (R^2 = 0.9771). The strongest overall performance was obtained using all three predictors, yielding a test R^2 of 0.9832, RMSE of 4.1551, MAE of 2.9724, and MAPE of 4.91%. In contrast, the combination of initial BG concentration and catalyst dosage without degradation time produced a negative test R^2 (-0.0873), indicating poor predictive performance relative to a mean-based baseline. These results broadly reinforce the SHAP findings: degradation time carries the strongest predictive signal, initial BG concentration provides substantial additional information, and catalyst dosage makes a smaller complementary contribution within the investigated $CaTiO_3$-assisted BG water-treatment system.

Table 3.

Comparative performance of the CatBoost model using different combinations of input features for predicting BG removal.

Feature combination	Data	R^2	MAE	RMSE	MSE	MedAE	MAPE
Degradation time	Training	0.688627	13.527166	15.473571	239.431408	14.570559	41.129861
	Testing	0.828717	10.560984	13.279174	176.336450	12.741073	20.017761
Degradation time + Catalyst dosage	Training	0.872172	6.490841	9.914306	98.293472	2.771041	21.268193
	Testing	0.900454	5.496901	10.123366	102.482538	0.612506	10.237318
Degradation time + Initial BG concentration + Catalyst dosage	Training	0.998496	0.804730	1.075454	1.156601	0.529218	2.133761
	Testing	0.983230	2.972424	4.155129	17.265099	1.773025	4.908494

In summary, the feature-combination results support the SHAP interpretation, which is that degradation time is the most important feature. provides the core predictive information; initial BG concentration adds a significant amount of predictive information. A complementary contribution, combined with the dosage of the catalyst, provides a smaller extra contribution. The three-variable model was thus included in the subsequent robustness analysis.

3.6. Noise robustness evaluation

A controlled amount of 'noise' (deceptive information) was added to the data sets, and the robustness of the selected CatBoost model was tested. Gaussian noise was added at 1, 2.5, 5, and 10% levels. As is evident in Fig. 8, the model continued to be very Stable under low to moderate levels of perturbation. The test R^2 only slightly deviated from the original value (0.9884) at 1% noise to 0.9883 at 2.5% and 0.9868 at 5%, while the corresponding RMSE values were 3.4629, 3.4674, and 3.6670, respectively. Even at the highest noise level of 10%, CatBoost retained a test R^2 of 0.9778, with RMSE and MAE increasing to 4.6611 and 3.4915. The slow rate of breakdown does not imply that the model was overly complicated. The system is sensitive to small changes in experimental variables. This is a strength. The assessment is especially applicable to datasets of literature-based photocatalytic data, Experimental variation may occur due to difference in measurement conditions and data quality.

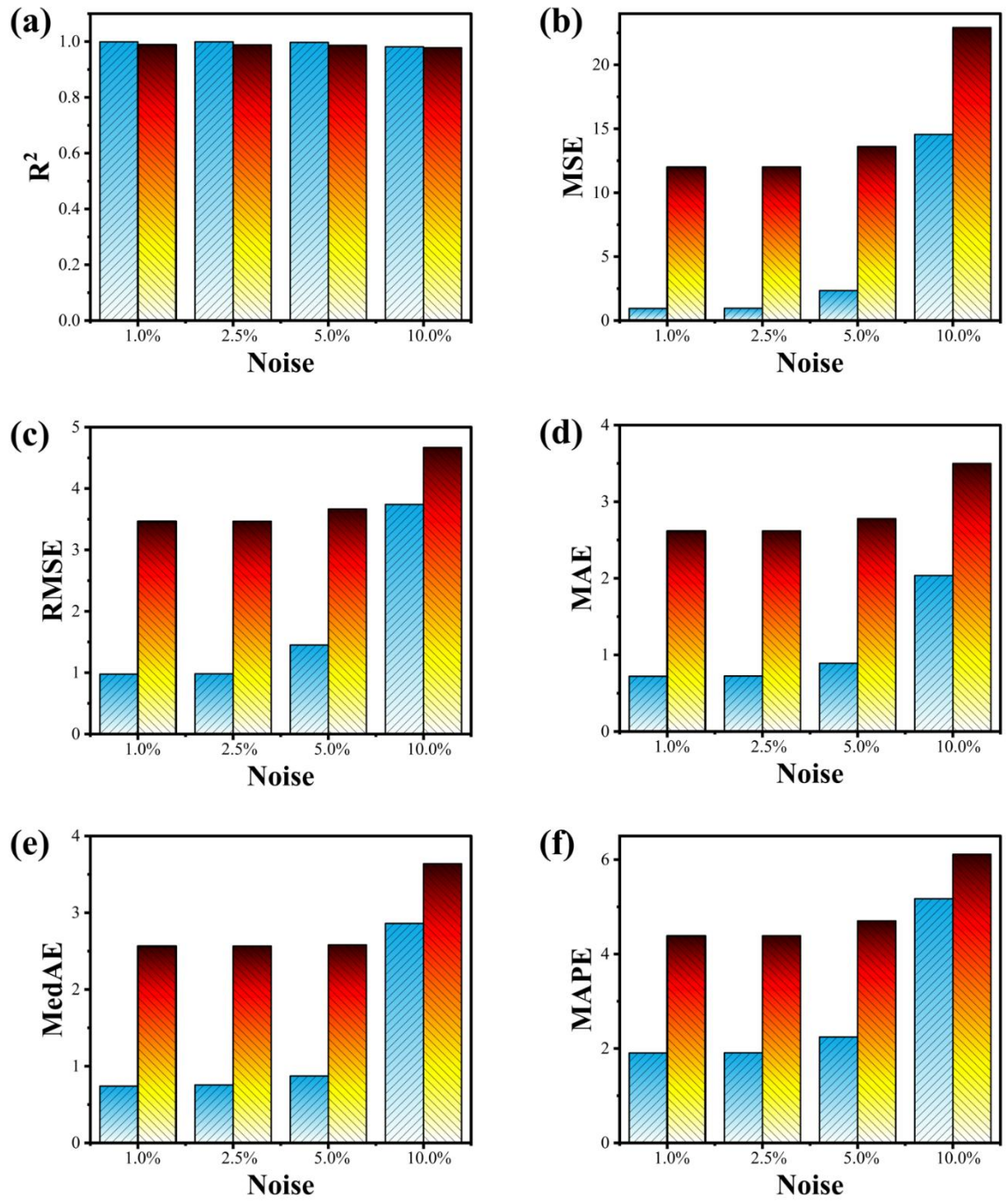


Fig. 8. The bar plots showing the comparison of training and testing metrics: (a) R^2, (b) MSE, (c) RMSE, (d) MAE, (e) MedAE, and (f) MAPE of the CatBoost model.

The selected experimental condition was then assessed using the final CatBoost prediction. At a catalyst dosage of 200 mg/L, an initial BG concentration of 10 mg/L, and a degradation time of 105 min, the model predicted a BG removal of 89.3154%, compared with the experimental value of 88.4615%. The corresponding prediction error was only 0.9652%, indicating close agreement between the model estimate and the reported experimental response. Recent

photocatalytic ML studies have likewise shown that CatBoost can provide accurate optimization-oriented predictions when nonlinear operating variables are modeled jointly.

Table 4.

Optimal experimental and predicted BG removal (%) values for the selected CatBoost model.

Catalyst dosage (mg/L)	Initial BG concentration (mg/L)	Degradation time (min)	Predicted BG removal (%)	Experimental BG removal (%)	Error (%)
200	10	105	89.3154	88.4615	0.9652

Overall, the limited loss of predictive accuracy under increasing noise, together with the close agreement at the selected experimental condition, supports the practical stability of the CatBoost model within the investigated data range.

3.7. Comparison with previous literature

The predictive performance obtained in the present study compares favourably with recent machine-learning applications in photocatalytic dye degradation, although direct numerical comparison should be made cautiously because the datasets, photocatalysts, dyes, and experimental ranges differ considerably among studies. As summarized in Table 5, recent work has increasingly favoured ensemble and boosting algorithms because they can capture nonlinear interactions among reaction variables more effectively than simpler single-model approaches. In a methylene-blue degradation study using $CuWO_4@TiO_2$, HistGradientBoosting achieved a test R^2 of 0.9915, while CatBoost and Gradient Boosting also performed strongly [52]. A separate study on Rhodamine B degradation using doped ZnO reported CatBoost as the best-performing model with an R^2 of 0.96 and further confirmed the predicted optimum experimentally [53].

Table 5.

Comparison of the present work with selected machine-learning studies on photocatalytic dye degradation.

Study	Photocatalytic system / pollutant	Main ML model	Reported performance
Jaffari et al.	$BiFeO_3$-based photocatalysts/ Malachite Green	CatBoost	CatBoost identified as the best-performing model [54].
Ahmed et al.	$CuWO_4@TiO_2$/ Methylene Blue	HistGradientBoosting	Test $R^2 = 0.9915$ [55].

ML-assisted ZnO study	Doped ZnO/ Rhodamine B	CatBoost	$R^2 = 0.96$ [56].
Nanozeolite study	Nanozeolite / Rhodamine B	ANN	$R^2 \approx 0.90$ on testing [57].
Present study	$CaTiO_3$ (2:7) / Brilliant Green	CatBoost	Test R^2 = 0.9884; repeated-CV R^2 = 0.9697

The present CatBoost model therefore falls within the performance range reported for recent high-performing photocatalytic ML studies. Its test R^2 of 0.9884, together with strong repeated-validation performance and low prediction error, is particularly encouraging given the comparatively small dataset used here. The result also agrees with the broader trend in the literature that boosted tree models frequently perform well for photocatalytic systems containing nonlinear and interacting operating variables. In a recent study of machine learning in photocatalytic degradation, too, the importance of interpretable models was stressed, as well as good validation when using ML for predicting dye removal, and standardized datasets [58].

However, there are also limitations of the present analysis. The model was derived from the following: used published experimental data for experiments instead of creating a large-scale database, and the final Only the $CaTiO_3$ 2:7 composition was allowed to be used for modelling space, as it was the composition that gave meaningful Changes in all three predictors. Thus, the model should be understood within the context of the ranges were based on the original experiments instead of extrapolation beyond them was free. Future studies should thus be extended to cover other catalyst compositions. The concentrations of pollutants, conditions of irradiation, and independent experimental batches. It would be good to validate it with data collected for the first time, to determine the ability of the model to transfer to other data sets. The current idea of such expansion is that the framework could become more than a predictive tool and can be extended to a more comprehensive decision support. The process of photocatalytic BG treatment is suggested.

4. Conclusion

This study developed an explainable machine-learning framework for predicting Brilliant Green (BG) removal from water through $CaTiO_3$-assisted photocatalysis. Using 63 published experimental observations for the 2:7 precursor composition, six regression algorithms were evaluated through Bayesian hyperparameter optimization and repeated cross-validation. Although Extra Trees achieved marginally higher holdout accuracy, CatBoost exhibited stronger repeated-validation performance, with a test R^2 of 0.9884, RMSE of 3.4629 percentage points, and mean cross-validation R^2 of 0.9697. SHAP analysis identified

degradation time (60.57%) as the dominant predictor, followed by initial BG concentration (32.12%) and catalyst dosage (7.31%). Feature-combination analysis reinforced these findings, while CatBoost maintained a test R^2 of 0.9778 under 10% Gaussian noise. At 200 mg/L catalyst dosage, 10 mg/L initial BG concentration, and 105 min irradiation, predicted removal (89.32%) closely matched the reported experimental value (88.46%). From a water-treatment perspective, these findings provide a practical basis for identifying influential operating variables, estimating removal efficiency within experimentally investigated conditions, and prioritizing promising treatment settings for subsequent laboratory verification. Such predictions could help researchers target experiments more efficiently rather than repeatedly screening every parameter combination. Nevertheless, the limited dataset and absence of independent validation constrain broader applicability. Future work should extend the framework to additional catalyst compositions, real wastewater matrices, and independently collected experiments, while assessing degradation intermediates, residual toxicity, and treatment energy requirements. These advances are essential for translating predictive insights into environmentally meaningful and operationally reliable photocatalytic water-treatment strategies.

CRediT Authorship Contribution Statement

Md. Abdul Hamid Khan: Conceptualization, Methodology, Data curation, Formal analysis, Software, Visualization, Investigation, Writing-original draft, Writing-review & editing. **Md. Mujahidul Islam:** Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Visualization, Writing-review & editing. **Shohanur Rahman:** Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Visualization, Writing-review & editing.

Acknowledgements

The authors gratefully acknowledge Ernawati et al. for making their experimental data on $CaTiO_3$-assisted Brilliant Green degradation publicly available, thereby enabling the present computational investigation. The authors also acknowledge the institutional resources and support provided by their affiliated institution.

Ethical Approval

Ethical approval was not required, as the study involved secondary analysis of previously published experimental data and did not include human participants, animals, personal information, or newly collected biological samples.

Consent to Participate

Not applicable. No human participants were involved in this study.

Consent for Publication

Not applicable. The manuscript does not contain identifiable personal information requiring consent for publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this manuscript.

Data Availability

The experimental data analysed in this study were reconstructed from the publicly available dataset reported by Ernawati et al. (2020): https://doi.org/10.1016/j.dib.2020.106099. The reconstructed dataset and supporting computational files are provided in the Supplementary Material accompanying this article.

Funding

This research received no external funding.

Declaration of Generative AI and AI-Assisted Technologies

The authors acknowledge the use of ChatGPT (OpenAI) and Grammarly to assist with language refinement, manuscript organization, and editorial improvements. ChatGPT was also used to support the organization and structuring of information extracted from the published experimental source. All extracted data were reviewed against the original publication, and the authors retained responsibility for the study design, computational analyses, interpretation of results, scientific conclusions, and accuracy of the manuscript. The AI-assisted tools were not treated as authors or independent sources of scientific evidence.

References:

1. Dutta, S. *et al.* Contamination of textile dyes in aquatic environment: Adverse impacts on aquatic ecosystem and human health, and its management using bioremediation. *J Environ Manage* **353**, 120103 (2024).
2. Environmental impacts and remediation of dye-containing wastewater | Nature Reviews Earth & Environment. https://www.nature.com/articles/s43017-023-00489-8.
3. Sustainable Solutions: Reviewing the Future of Textile Dye Contaminant Removal with Emerging Biological Treatments. https://www.mdpi.com/2300-7575/24/2/7.
4. Kavitha, G., Govindhan, M. & Premkumar, S. Dye pollution and its implications for human health, aquatic ecosystems, and sustainable wastewater treatment: A comprehensive review. *Journal of Water Process Engineering* **80**, 109071 (2025).
5. Singh, G. B., Vinayak, A., Mudgal, G. & Kesari, K. K. Azo dye bioremediation: An interdisciplinary path to sustainable fashion. *Environmental Technology & Innovation* **36**, 103832 (2024).
6. Aragaw, T. A. A review of dye biodegradation in textile wastewater, challenges due to wastewater characteristics, and the potential of alkaliphiles. *Journal of Hazardous Materials Advances* **16**, 100493 (2024).
7. El Messaoudi, N. *et al.* Advancements in adsorption and photocatalytic degradation technologies of brilliant green from water: Current status, challenges, and future prospects. *Materials Today Chemistry* **42**, 102399 (2024).
8. Kumari, N., Chintakula, S., Sai Sonali Anantha, I. & Maddila, S. An efficient P3TA/Fe doped TiO2 catalyst for photo-degradation of Brilliant green dye and inactivation of pathogens under visible light. *Results in Chemistry* **5**, 100759 (2023).

9. Chen, C.-C. *et al.* Photocatalyzed *N*-de-ethylation and degradation of Brilliant Green in TiO2 dispersions under UV irradiation. *Desalination* **219**, 89–100 (2008).
10. Khandelwal, D., Rana, I., Mishra, V., Ranjan, K. R. & Singh, P. Unveiling the impact of dyes on aquatic ecosystems through zebrafish – A comprehensive review. *Environmental Research* **261**, 119684 (2024).
11. Dutta, S. *et al.* Contamination of textile dyes in aquatic environment: Adverse impacts on aquatic ecosystem and human health, and its management using bioremediation. *Journal of Environmental Management* **353**, 120103 (2024).
12. Khandelwal, D., Rana, I., Mishra, V., Ranjan, K. R. & Singh, P. Unveiling the impact of dyes on aquatic ecosystems through zebrafish – A comprehensive review. *Environmental Research* **261**, 119684 (2024).
13. Kavitha, G., Govindhan, M. & Premkumar, S. Dye pollution and its implications for human health, aquatic ecosystems, and sustainable wastewater treatment: A comprehensive review. *Journal of Water Process Engineering* **80**, 109071 (2025).
14. Patel, D., Singh, A., Ambati, S. R., Singh, R. S. & Sonwani, R. K. An overview of recent advances in treatment of complex dye-containing wastewater and its techno-economic assessment. *Journal of Environmental Management* **370**, 122804 (2024).
15. Aragaw, T. A. A review of dye biodegradation in textile wastewater, challenges due to wastewater characteristics, and the potential of alkaliphiles. *Journal of Hazardous Materials Advances* **16**, 100493 (2024).
16. A Thorough Examination of Photocatalysis for Addressing Wastewater Pollution - Islam - 2025 - ChemistrySelect - Wiley Online Library. https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/slct.202500274.

17. A review on machine learning in photocatalytic degradation of organic dyes from wastewater: Current trends and future directions - ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S2214714425018331.
18. Machine Learning-Enabled Photocatalytic Wastewater Treatment: Recent Advances in Catalyst Design, Performance Prediction, and Process Optimization. https://www.mdpi.com/2073-4344/16/9/786.
19. Chen, C.-C. *et al.* Photocatalyzed *N*-de-ethylation and degradation of Brilliant Green in TiO2 dispersions under UV irradiation. *Desalination* **219**, 89–100 (2008).
20. Photocatalytic effect of TiO2and the effect of dopants on degradation of brilliant green | Sustainable Chemical Processes | Springer Nature Link. https://link.springer.com/article/10.1186/2043-7129-1-4.
21. Sood, S., Umar, A., Kumar Mehta, S., Sinha, A. S. K. & Kumar Kansal, S. Efficient photocatalytic degradation of brilliant green using Sr-doped TiO2 nanoparticles. *Ceramics International* **41**, 3533–3540 (2015).
22. Removal of brilliant green dye by synergistic effect of adsorption and photocatalysis on B-TiO2 | Environmental Science and Pollution Research | Springer Nature Link. https://link.springer.com/article/10.1007/s11356-026-37915-5.
23. Visible light driven degradation of brilliant green dye using titanium based ternary metal oxide photocatalyst - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2211379719302475.
24. Kumari, N., Chintakula, S., Sai Sonali Anantha, I. & Maddila, S. An efficient P3TA/Fe doped TiO2 catalyst for photo-degradation of Brilliant green dye and inactivation of pathogens under visible light. *Results in Chemistry* **5**, 100759 (2023).

25. Vinodha, S., Suganya, M., Shanmugam, N., Senthil Kumar, A. & Rajarajan, M. Optimization of the dosage of ZnO photocatalyst for the degradation of sunlight driven multi dose organic dye brilliant green. *Journal of the Indian Chemical Society* **101**, 101328 (2024).
26. Gnanaprakasam, A., Sivakumar, V. M. & Thirumarimurugan, M. A study on Cu and Ag doped ZnO nanoparticles for the photocatalytic degradation of brilliant green dye: synthesis and characterization. *Water Sci Technol* **74**, 1426–1435 (2016).
27. Khalid, A., Ahmad, P. & Rafiq, M. Unveiling the synthesis of visible-light-induced indium oxide supported graphitic carbon nitride (In2O3/g-C3N4) for photocatalytic degradation of brilliant green dye. *Polyhedron* **284**, 117907 (2026).
28. El Messaoudi, N. *et al.* Advancements in adsorption and photocatalytic degradation technologies of brilliant green from water: Current status, challenges, and future prospects. *Materials Today Chemistry* **42**, 102399 (2024).
29. Nuzhat, T., Radzi, N. A. B. M., Maruthupandy, M. & Yussof, S. A review on machine learning in photocatalytic degradation of organic dyes from wastewater: Current trends and future directions. *Journal of Water Process Engineering* **78**, 108760 (2025).
30. Ernawati, L. *et al.* Experimental data of CaTiO3 photocatalyst for degradation of organic pollutants (Brilliant green dye) – Green synthesis, characterization and kinetic study. *Data in Brief* **32**, 106099 (2020).
31. Artificial Intelligence-Guided Supervised Learning Models for Photocatalysis in Wastewater Treatment. https://www.mdpi.com/2673-2688/6/10/258.

32. Du, Y. *et al.* Harnessing machine learning for energy optimization and intelligent process control in wastewater treatment. *Journal of Water Process Engineering* **80**, 109210 (2025).
33. Machine Learning-Enabled Photocatalytic Wastewater Treatment: Recent Advances in Catalyst Design, Performance Prediction, and Process Optimization. https://www.mdpi.com/2073-4344/16/9/786.
34. Nuzhat, T., Radzi, N. A. B. M., Maruthupandy, M. & Yussof, S. A review on machine learning in photocatalytic degradation of organic dyes from wastewater: Current trends and future directions. *Journal of Water Process Engineering* **78**, 108760 (2025).
35. Ernawati, L. *et al.* Experimental data of CaTiO3 photocatalyst for degradation of organic pollutants (Brilliant green dye) – Green synthesis, characterization and kinetic study. *Data in Brief* **32**, 106099 (2020).
36. XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://dl.acm.org/doi/10.1145/2939672.2939785.
37. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. in *Advances in Neural Information Processing Systems* vol. 31 (Curran Associates, Inc., 2018).
38. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Preprint at https://doi.org/10.48550/arXiv.1705.07874 (2017).
39. Kim, C.-M., Jaffari, Z. H., Abbas, A., Chowdhury, M. F. & Cho, K. H. Machine learning analysis to interpret the effect of the photocatalytic reaction rate constant (k) of

semiconductor-based photocatalysts on dye removal. *Journal of Hazardous Materials* **465**, 132995 (2024).

40. Metal oxide-based photocatalysts for the efficient degradation of organic pollutants for a sustainable environment: a review | Nanoscale Advances | The Royal Society of Chemistry. https://pubs.rsc.org/na/article/6/19/4781/856100/Metal-oxide-based-photocatalysts-for-the-efficient.
41. Ernawati, L. *et al.* Experimental data of CaTiO3 photocatalyst for degradation of organic pollutants (Brilliant green dye) – Green synthesis, characterization and kinetic study. *Data in Brief* **32**, 106099 (2020).
42. Ali, H. *et al.* Machine learning approach for photocatalysis: An experimentally validated case study of photocatalytic dye degradation. *Journal of Environmental Management* **386**, 125683 (2025).
43. Scyphers, M. E., Missik, J. E. C., Kujawa, H., Paulson, J. A. & Bohrer, G. Bayesian Optimization for Anything (BOA): An open-source framework for accessible, user-friendly Bayesian optimization. *Environmental Modelling & Software* **182**, 106191 (2024).
44. Jaffari, Z. H. *et al.* Machine learning approaches to predict the photocatalytic performance of bismuth ferrite-based materials in the removal of malachite green. *Journal of Hazardous Materials* **442**, 130031 (2023).
45. Ahmed, Y. *et al.* Advanced ciprofloxacin quantification: A machine learning and metaheuristic approach using ultrasensitive chitosan-gold nanoparticle based electrochemical sensor. *Journal of Environmental Chemical Engineering* **13**, 115094 (2025).

46. Steen, B., Broennimann, O., Maiorano, L. & Guisan, A. How sensitive are species distribution models to different background point selection strategies? A test with species at various equilibrium levels. *Ecological Modelling* **493**, 110754 (2024).
47. Bakır, R., Orak, C. & Horoz, S. Enhancing photocatalytic degradation of hazardous pollutants with green-synthesized catalysts: A machine learning approach. *Journal of Environmental Management* **385**, 125695 (2025).
48. Sohrabi, S. *et al.* Evaluation of machine learning and molecular dynamics models for photocatalytic water decontamination. *Process Safety and Environmental Protection* **195**, 106780 (2025).
49. K c, A., Rao, C. S. & Nair, V. Combination of ensemble machine learning models in photocatalytic studies using nano TiO2 - Lignin based biochar. *Chemosphere* **352**, 141326 (2024).
50. Machine learning modeling to predict the photocatalytic degradation rate constant in wastewater using TiO2 | Scientific Reports. https://www.nature.com/articles/s41598-026-51108-7.
51. From local explanations to global understanding with explainable AI for trees | Nature Machine Intelligence. https://www.nature.com/articles/s42256-019-0138-9.
52. Ahmed, Y. *et al.* Optimizing photocatalytic dye degradation: A machine learning and metaheuristic approach for predicting methylene blue in contaminated water. *Results in Engineering* **25**, 103538 (2025).
53. Xu, K. *et al.* Machine learning-assisted ZnO-doped photocatalytic degradation of rhodamine B: Prediction, analysis, and optimization. *Environmental Technology & Innovation* **40**, 104425 (2025).

54. Jaffari, Z. H. *et al.* Machine learning approaches to predict the photocatalytic performance of bismuth ferrite-based materials in the removal of malachite green. *Journal of Hazardous Materials* **442**, 130031 (2023).
55. Ahmed, Y. *et al.* Optimizing photocatalytic dye degradation: A machine learning and metaheuristic approach for predicting methylene blue in contaminated water. *Results in Engineering* **25**, 103538 (2025).
56. Xu, K. *et al.* Machine learning-assisted ZnO-doped photocatalytic degradation of rhodamine B: Prediction, analysis, and optimization. *Environmental Technology & Innovation* **40**, 104425 (2025).
57. Oviedo, L. R., Druzian, D. M., Nora, L. D. D. & da Silva, W. L. Study of machine learning on the photocatalytic activity of a novel nanozeolite for the application in the Rhodamine B dye degradation. *Catalysis Today* **443**, 114986 (2025).
58. Oviedo, L. R., Druzian, D. M., Nora, L. D. D. & da Silva, W. L. Study of machine learning on the photocatalytic activity of a novel nanozeolite for the application in the Rhodamine B dye degradation. *Catalysis Today* **443**, 114986 (2025).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx