Machine learning approach for photocatalysis: An experimentally validated case study of photocatalytic dye degradation

Hassan Ali[1], Muhammad Yasir[1], Hamza Ul Haq[2], Ali Can Guler[1], Milan Masar[1], Muhammad Nouman Aslam Khan[2,*], Michal Machovsky[1,*], Vladimir Sedlarik[1], Ivo Kuritka[1]

[1]Centre of Polymer Systems, Tomas Bata University in Zlin, Tr. T. Bati 5678, 76001 Zlin, Czech Republic

[2]Laboratory of Alternative Fuel and Sustainability, School of Chemical and Materials Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan

*Corresponding authors. Tel: +420-57-603-1110. E-mail: machovsky@utb.cz, mnouman@scme.nust.edu.pk

Abstract

In this study, machine learning (ML) models coupled with genetic algorithm (GA) and particle swarm optimization (PSO) were applied to predict the relative influence of experimental parameters of photocatalytic dye removal. Specifically, the impact of bandgap, dye concentration, photocatalyst dosage, solution volume, specific surface area, and time duration on photocatalytic degradation rate constant of cationic dyes was discerned using selected ML models, i.e., ensembled learning tree (ELT), gaussian process regression (GPR), support vector machine (SVM), and decision tree (DT). Thus, the data points were sourced from literature studies recently published in 2024 and 2023 on materials related to working on fundamental principles of photocatalysis. The ELT-PSO hybrid model outperformed all models with R^2 = 0.992 and RMSE = 2.6408e^{-04}, followed by DT, GPR, and SVM. The partial dependence plots and Shapley's analysis demonstrate that the type of dye, bandgap, dye initial concentration, and time duration are essential parameters for photocatalytic degradation, while sensitivity analysis further displayed solution volume and time duration to be the most influential parameters for rate constant determination. The optimized ML model's prediction was also experimentally validated using as-synthesized different compositions of Cu_2O/WO_3 heterostructures and ZnO nanoparticles. The results suggest that an ML-optimized study can be used in designing photocatalysts with optimum properties desired for the removal of cationic dyes at high rates from wastewater, thus saving energy and cost for a sustainable environment.

Keywords: machine learning; photocatalytic degradation; wastewater treatment; cationic dyes; optimization removal

1. Introduction

The remediation of environmental pollutants using photocatalysis has attracted significant interest over the past few decades, mainly due to simplified processes requiring only suitable photocatalysts and light irradiation (Rashid et al., 2024; Xiong et al., 2024). Photocatalytic degradation of organic contaminants, as well as other advanced applications such as N_2 and CO_2, proceed via a similar mechanism, i.e., production of photoexcited electrons (e^-), holes (h^+), and reactive oxygen species which participate collectively and react with the targeted molecular species to reduce or oxidize them into more benign and environmentally byproducts (Hua et al., 2024; L. Li et al., 2024; Luo et al., 2022). To date, various types of photocatalytic materials have been investigated; among them, metal oxides are the most common ones, such as TiO_2, ZnO, WO_3, and SnO_2, while other metal-free semiconductors like graphitic carbon nitride (g-C_3N_4) have also gained increased scientific interest owing to its visible light harvesting and suitable band-edge positions (Akram et al., 2024; Imai et al., 2024; Saeidpour et al., 2022). However, despite significant progress, a viable commercial system on a large scale for water treatment is yet to be realized as current photocatalytic efficiencies are still significantly lower due to contributing physiochemical properties of the catalyst (Ollis, 2018). Therefore, current efforts are focused on improving photocatalytic efficiency by reducing backward charge recombination and extending the light absorption spectrum (Akshhayya et al., 2023; Ali et al., 2023). In this regard, strategies like the formation of heterojunction, metal-doping, defect engineering, photo-Fenton process, and LSPR effect have been explored (J. Li et al., 2024; Sun et al., 2024). However, besides contriving photocatalysts based on preceding strategies, there are other operational parameters that affect the rate of photocatalytic degradation, such as bandgap, surface area, initial volume, pollutant concentration, and time duration of degradation (Gusain et al., 2020a). For example, Shana et. al. study revealed that the photocatalytic degradation of crystal violet (CV), a known carcinogen is dependent on various factors conditions including contact time, photocatalyst dosage, and bandgap (Shabna et al., 2024). These factors, along with others, collectively determine the yield of photocatalytic reaction efficiency and rate constant (k). However, it is difficult to analyze the relative influence of each factor in comparison to other factors as photocatalytic degradation proceeds by multi-variable interactions, the determination

of which can potentially enable viable large scale photocatalytic reactors for wastewater treatment (Sacco et al., 2020). Understanding the impact of each factor on photocatalytic efficiency is crucial for developing large-scale systems where energy and cost become imperative for commercial viability. Considering that most literature studies on photocatalytic materials omit optimization of operational factors while solely reporting on selected parameters, determining the inter-relation of such parameters can provide useful insights.

Fortunately, with recent advances in computer science and machine learning (ML), complex relationships between various factors can be determined by harnessing the power of ML (Baskar et al., 2024; Freiesleben et al., 2024). ML is a subfield of artificial intelligence in which different types of mathematical algorithms are used to analyze patterns in a specific dataset and to draw predictions and correlations (Anjaneyulu et al., 2024; Bakır et al., 2024). Specifically, obtaining reasonable predictions via ML requires the acquisition of a large dataset, applying appropriate algorithms based on the dataset, model training, and finally, model validation to establish its authenticity (Lemm et al., 2023). ML has been successfully implemented in various industries and applications to make predictions and draw conclusions from specific datasets (Banisheikholeslami and Qaderi, 2024; Hancock et al., 2023; Knott et al., 2023; Weinreich et al., 2023). In this context, various types of ML algorithms have been investigated, with each having its suitability, applicability, and versatility in making predictions and drawing conclusions, such as linear regression, K-Nearest neighbors (KNN), support vector machines (SVM), decision tree (DT), generative model, and random forest (Breiman, 2001; Long et al., 2024; Quinlan, 1986). As with other research domains, literature survey reveals that ML has been recently implemented to investigate various aspects of photocatalysis and design of materials. For instance, Chen et. al. utilized ML-assisted approach for optimized designing of $BiWO_6$/MIL-53(Al) nanocomposite photocatalyst possessing optimized removal of Rhodamine Blue (RhB) (Zhai and Chen, 2023). By considering 13 nanocomposite synthesis conditions, the authors gathered 55 points dataset and utilized various ML models to ascertain the role of each feature which enabled ML-assisted design of $BiWO_6$/MIL-53(Al) nanocomposite. Similarly, Jiang group's developed an artificial neural network for predicting photocatalytic rate constant of organic contaminants using literature sourced dataset consisting of six metal oxide photocatalysts (Jiang et al., 2021). Their developed model was able to predict photocatalytic degradation rate constant within considerable experimental values according to R^2 and RMSE values of 0.746 and 0.293, respectively. Other

authors also utilized ML-assisted models to predict the features influencing the rate constant of the photocatalytic reactions. Recently, Kim et. al conducted a thorough study on five different semiconductors for prediction of photocatalytic degradation rate constant of malachite green (MG) and indigo dyes (Kim et al., 2024). Via controlled experimental conditions, the authors were able to predict rate constant using dataset sourced from their own experiments which also revealed the influence of various experimental parameters according to SHAP analysis. According to their dataset comprising of five different photocatalysts, Ni%, initial concentration of dye, catalyst dosage, reaction time were reported to be most influencing features for five different types of selected photocatalysts. However, the preceding studies incorporated photocatalytic materials on a limited scale with few different types of photocatalysts. Considering that very few studies have been conducted on establishing the relative influence of selected operational parameters on photocatalytic rate constant, we decided to extend the contemporary knowledge by applying ML on a large dataset obtained from the literature survey to elucidate the relative influence of parameters. To leverage the emerging ML technology, we surveyed numerous reports on the photocatalytic removal of dyes, specifically, methylene blue (MB), CV, RhB, and MG and collected over 300 data points for the ML dataset. In this study, we applied different ML algorithms, i.e., ensembled learning tree (ELT), gaussian process regression (GPR), SVM, and DT, coupled with genetic algorithm (GA) and particle swarm optimization (PSO), to establish the intricate relative influence of photocatalytic dye degradation parameters commonly encountered in laboratory settings. In contrast to other reported studies, herein, we have integrated the ML algorithm with GA and PSO, which enables more accurate and feature-rich prediction of various factors, as shown in **Figure 1**.

The contribution and novelty of our current work are as follows: (i) semiconductor-based photocatalysts without secondary functionalization, such adsorbents work on the fundamental principles of photocatalysis and have been utilized for the treatment of synthetic dyes. We have utilized recently reported studies to obtain a dataset encompassing more than 65 different types of photocatalysts for cationic dye treatment, particularly MB. (ii) according to our knowledge, currently, no study has been conducted on utilizing the selected ML models together with metaheuristics-based optimization on such a large dataset consisting of a variety of different photocatalyst types, and (iii) the developed ML models were not only validated experimentally

using testing dataset sourced from the literature reports, but also laboratory synthesized two different photocatalysts, i.e., ZnO and WO_3.

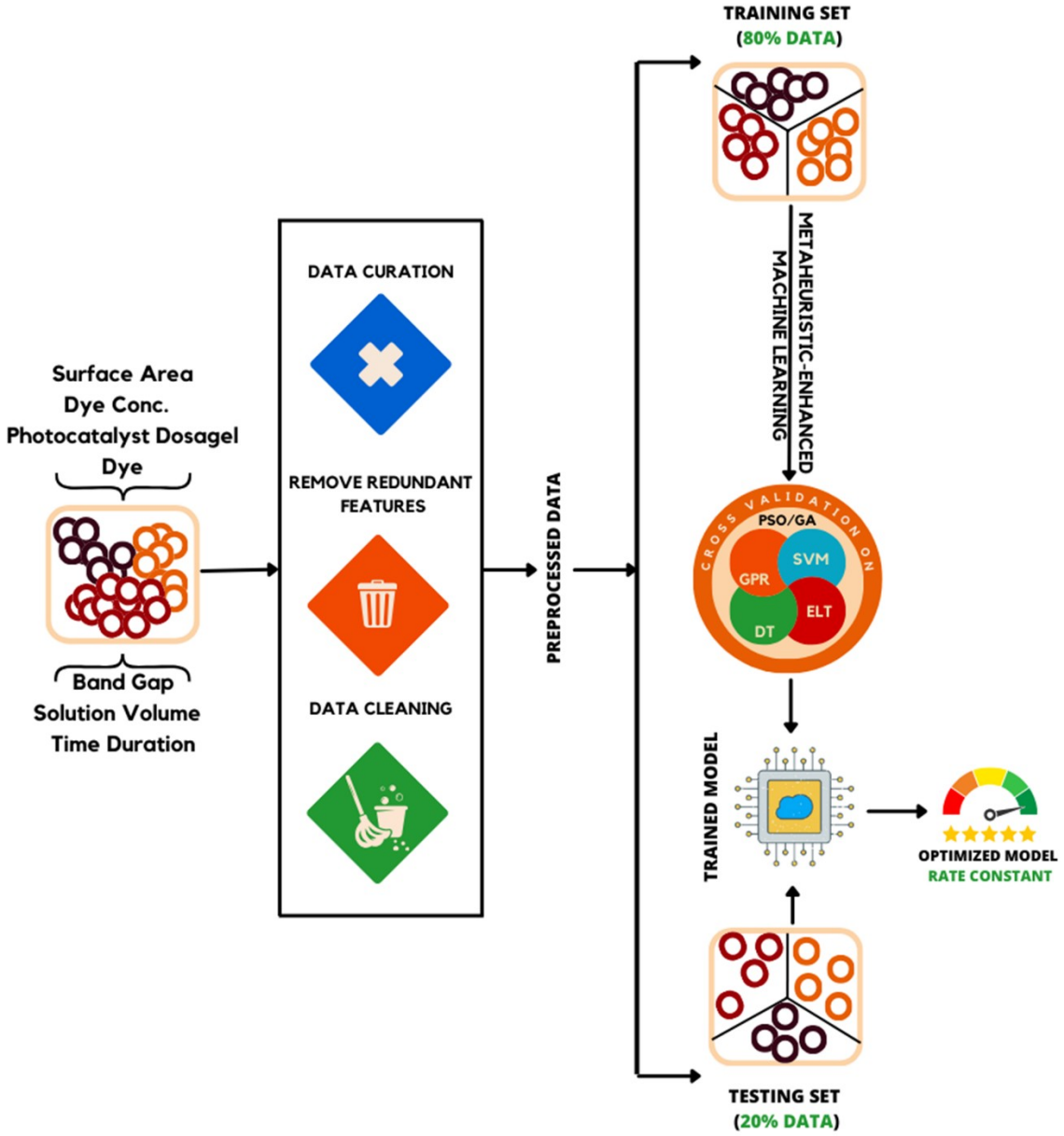

Figure 1. Schematic demonstration of machine learning methodology workflow.

2. Materials and Methods

2.1 Materials and reagents

For the validation of the developed model prediction, ZnO photocatalyst was prepared using zinc acetate dihydrate, p.a., (Penta, Czech Republic) and oxalic acid dihydrate, p.a., (Penta, Czech Republic) as starting precursors. An extended description of the preparation method is reported in our previous literature study (Sedlak et al., 2020).

2.2 Data collection, representation, and pre-processing

ML requires a large dataset, which is parsed through the appropriate models for subsequent training and making relevant predictions (Taye, 2023). Therefore, in this study, we collected data points from various reported studies on photocatalytic dye removal using the Web of Science (WoS) index using the search keyword, photocatal* in the title, and filtering out the obtained results by using additional “methylene blue” (MB), a common cationic dye to keep consistency. In some of the considered studies, besides MB, photocatalytic degradation was also conducted on other similar dyes, which were also included in the dataset. Experimental results have shown that cationic dyes exhibit a similar rate of photodegradation owing to their chemical structure (Verma et al., 2021). Therefore, we opted to include only cationic dyes to keep the dataset consistent. The dataset obtained from WoS is provided in **Table S1**. Moreover, data were extracted only from studies reliant on the fundamental principles of photocatalysis, excluding any major functionalization unrelated to photocatalysis, such as the Fenton process and highly adsorbent non-photocatalytic materials. The following independent variables were extracted from the literature reports: bandgap (1.02-3.90), dye concentration (1-100 mg/L), photocatalyst dosage (2-1000 mg), solution volume (10-1000 mL), specific surface area (4.5-651 m^2/g), and time duration (20-240 min), while photocatalytic reaction rate constant was taken as the dependent variable. The ranges of experimental input variables are presented in **Figure 2**. The data distribution of the sourced dataset is also presented visually in the form of count plot and KDE histograms, box-whisker, and violin plots (**Figure S1)**, while the tabulated form of data distribution, skewness, kurtosis, mean, and standard deviation stats are also given in **Table S2**. It should be noted that in contrast to sourcing datasets from controlled laboratory experiments, sourcing from literature-sourced studies is rather limited due to the lack of availability of all relevant features and parameters that can influence the photocatalytic process. Nevertheless, a general predictive inference from the features can still be determined by considering reliable

dataset features that are commonly reported and influence photocatalytic rate constant to a great extent. Moreover, it should be noted that some influencing features do not pose a major influence on the dependent variable as they were not varied. For instance, temperature can significantly influence the photocatalytic rate constant (Siddique et al., 2019), but it is assumed that photocatalytic experiments are carried out at room temperature unless stated otherwise, and thus, any influencing contribution of temperature from the dataset is excluded on the whole. Some studies reported thermodynamic studies on the photocatalytic dye degradation with temperature variation, however, such variations were not included in the dataset. After data acquisition, the data points were processed for normalization and removal of any detected outliers. Outliers were expected to occur in the obtained dataset as most studies were obtained in different laboratory settings and research groups. The dataset was then randomly divided into training sets (80%) and testing points (20%) and subsequently used for ML training and validation of applied selected models. To identify outliers, the 'quartiles' method was used since the given data in this study is nonlinear and not normally distributed. The total percentage of outliers in each variable is as follows: band gap (0.66 %), dye concentration (4.9%), photocatalyst dosage (11.5%), solution volume (9.2%), surface area (3.7%), time duration (0%), and rate constant (4.7%). To deal with outliers, different techniques, including moving average, least square model, cubic spline, and Savitzky Golay filter, were applied. Cubic spline reported a maximum 2-norm error of 0.015 in the input variable, photocatalyst dosage, which is minimal in comparison to other techniques.

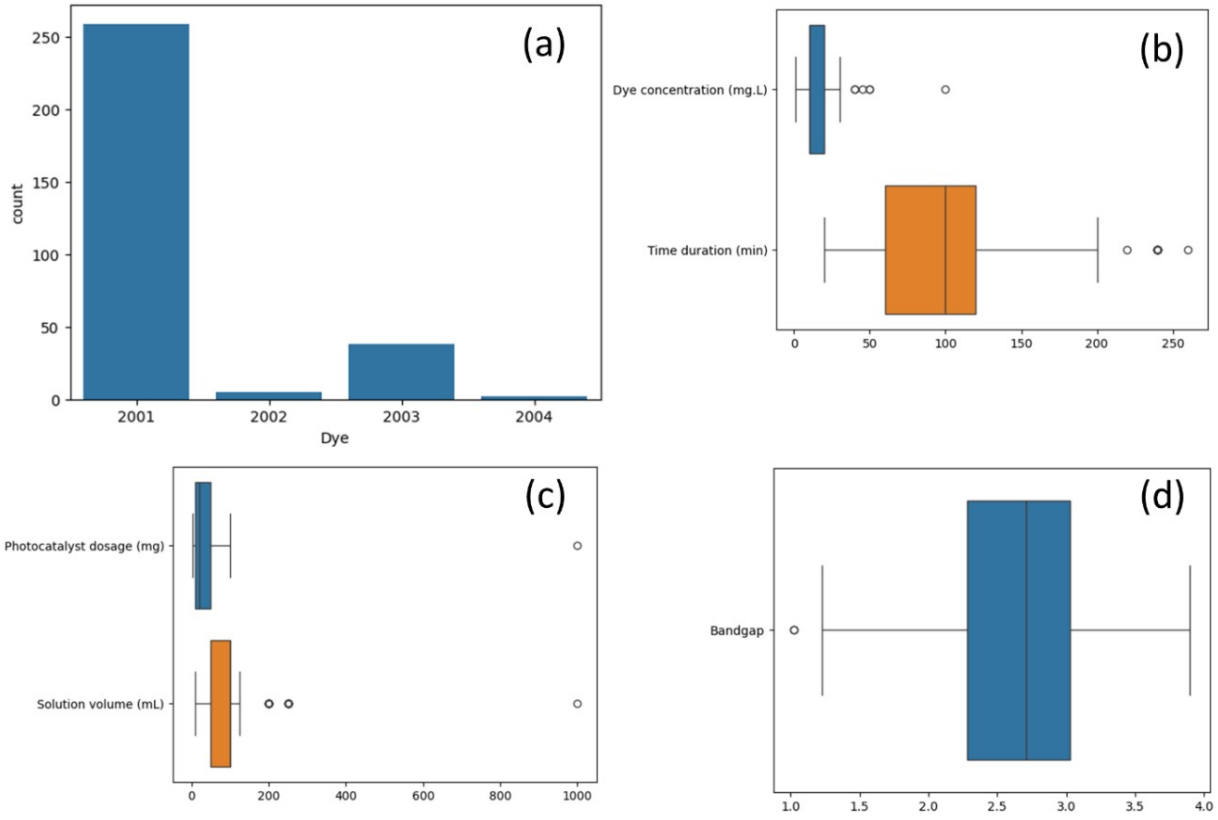

Figure 2. Count plot (a) and box-whisker plots (b-d) against the available dataset for four dyes (Designated no: 2001 = MB, 2002 = CV, 2003 = RhB, and 2004 = MG).

2.3 Machine learning models and optimization techniques

The evaluation metrics and description of SVM, DT, ELT, and GPR models are elaborated in sections 1.1 and 1.2 (supplementary data). The performance of evaluated models is further optimized using GA and PSO methods (supplementary sections 1.3 and 1.4). The selection of four ML models among various available ML models was made by taking into account complex and non-linear interactions among experimental variables influencing the photocatalytic reaction and literature studies on implemented ML models yielding excellent results for chemical science-based applications (Bang Truong et al., 2023; Gao et al., 2025; Li et al., 2023; Tu et al., 2023). ELT model is highly effective for capturing intricate patterns and enhancing predictive accuracy while mitigating overfitting through model averaging (K C et al., 2024). GPR was chosen as it is particularly suitable for small datasets, offering a probabilistic framework that models smooth non-linear functions and provides uncertainty quantification (Chow et al., 2022). SVMs, employing kernel functions, are capable of modeling complex relationships and exhibit strong generalization performance via the maximum-margin principle (Akyildiz et al., 2024). In the

case of DT, although a single DT typically yields lower predictive performance compared to the aforementioned models, it offers superior interpretability, enabling clear identification of dominant features and threshold-based decision rules (Özsoysal et al., 2024). In contrast to linear models, which lack the flexibility to capture non-linear dependencies, and neural networks, require large datasets and are less interpretable, the selected models provide a balanced approach for high predictive accuracy of photocatalytic dye degradation rate constant while maintaining interpretability and computational feasibility. Thus, the training dataset was uploaded to MATLAB's Regression Learner App with 5-fold cross-validation because MATLAB offers a variety of ML models, broadly categorized into different families. Within each category, Bayesian Optimization, a built-in optimization technique in MATLAB was employed to fine-tune the respective models. Then, the predictive capability of the best-performing model in each category was further enhanced through feature engineering and hyperparameter tuning. To further optimize the developed ML models, metaheuristic algorithms, GA and PSO optimization were done, which have shown a tremendous integration with ML models to enhance predictive capability using feature engineering and hyperparameter tuning (Khan et al., 2022). For PSO, the swarm size is 25 particles with a maximum limitation of 100 iterations and stall iterations of 50. The initial swarm matrix is randomly initialized with bounds of photocatalysis parameters, and the inertia range is between 0.1 and 1.1, with a self-adjustment weight of 1.49. In the case of GA, the population size is 50, with a maximum generation of 100 and a stall generation of 25. A scattered crossover technique is used, and Gaussian mutation is applied with a dynamic mutation rate and elite count of 2.

3. Results and discussion

3.1 Pearson correlation heatmap

To establish a general linear correlation between the selected variable features used in this study, a Pearson correlation heatmap was obtained, as shown in **Figure 3**. Interpretation of the data presented is intuitive, as the relative gradient of color indicates positive or negative among the selected features as highlighted by the scale shown on the right side of correlation heatmap (Zheng and Wu, 2019). Numerically, a value closer to one implies a stronger correlation and vice versa. From the heatmap, it can be observed that there is no significant correlation observed for pollutants, which is in accordance with the dataset, i.e., only photocatalytic experiments conducted for degradation of cation synthetic dyes were selected; the majority of those studies

were on MB. Therefore, changing cationic dyes such as RhB are less likely to influence changes in external parameters. In the case of bandgap, a negative correlation is observed for rate constant, which is as expected since photocatalytic materials with high bandgap have low capacity to absorb the full spectrum of light and, therefore, yield low production of excited reactive species to derive practical photocatalytic applications. For example, photocatalysts capable of harvesting solar irradiation in only UV region cannot absorb the visible light spectrum which consists of around 40 % of the solar spectrum leading to reduced efficiency (Johar et al., 2015). Moreover, the bandgap has a negligible relation (-0.03) with dye because all the dyes considered in the modeling are of the same type (cationic). In the case of dye concentration, the strongest correlation is observed for photocatalyst dosage and solution volume, respectively. Here, it should be noted that a stronger correlation (+0.59) between dye conc. and photocatalyst dosage is observed rather than between dye conc. and solution volution. This is because experiments were conducted with a high dosage of photocatalyst at higher concentrations, as expected, providing a higher correlation. Similarly, photocatalyst dosage and solution volume are observed to have a high correlation with a + 0.73 value, which could be due to the reason that higher amounts of solution volumes require a higher dosage of photocatalyst to produce a similar degradation rate, which has a linear relationship. Although very few studies included in the dataset utilized in this study provided or evaluated surface area for the study, we still included this parameter in order to increase the predictive evaluation of the ML models. However, due to the limited provision of surface area data in reported studies, a highly accurate correlation between surface area and selected parameters cannot be established. The rate of time duration indicates a negative correlation with the rate constant, implying that a higher yield of the rate constant is observed at a low time duration; as expected, a high rate of removal activity is observed in the initial phase of the experimental run with gradually achieving a stable rate as dye molecules are adsorbed and surface sites are saturated.

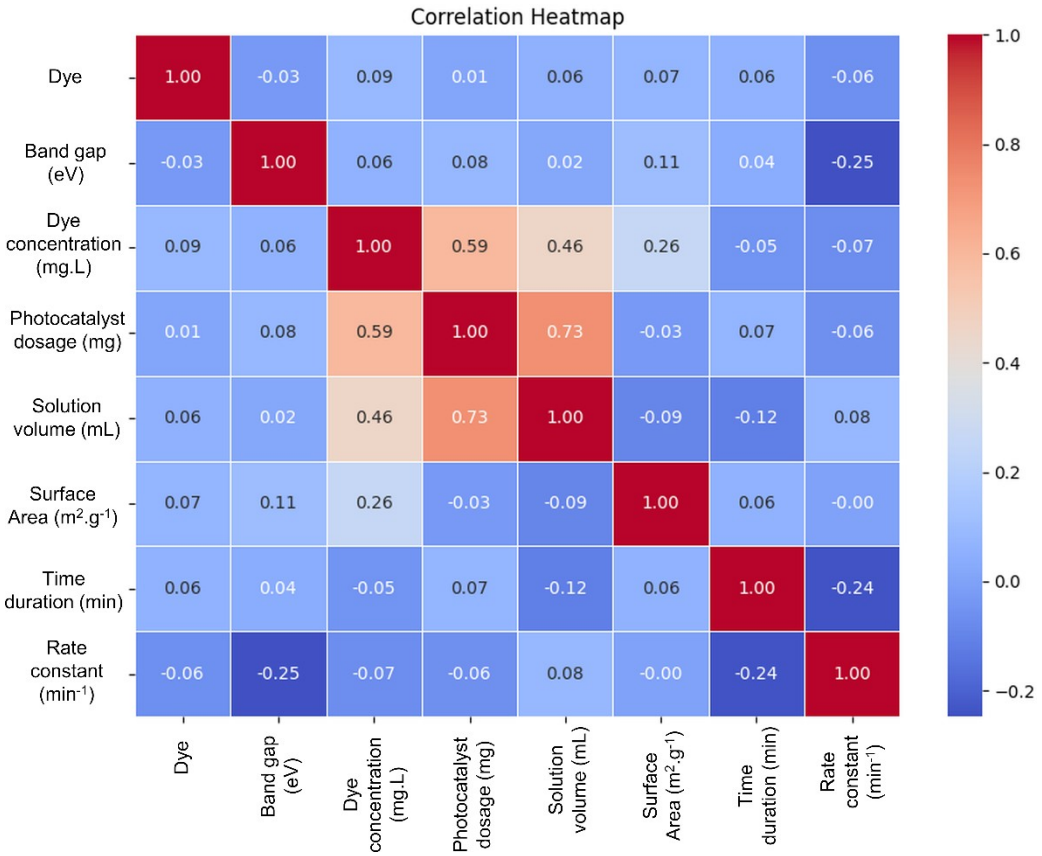

Figure 3. Pearson correlation between the input parameters (experimental conditions and photocatalyst properties) and output (rate constant).

3.2 Models feature selection and hyperparameter tuning

The most influencing features after fine-tuning of ML models were selected and further used to optimize the models using GA and PSO algorithms. For the DT model, all features were selected except dye concentration for GA, while dye and solution volume were excluded, considering the rest of the parameters, which were major predictors for the model. For the ELT model, under the GA algorithm, all input parameters except dye conc., photocatalyst dosage, and solution volume were included in the model, while the PSO method considered all the input parameters for model prediction optimization. For optimization of the GPR ML model, all factors except solution volume and surface area were considered for GA optimization, while in the PSO algorithm, only input parameters, surface area, and time duration were not considered. In contrast to other models, the input parameters for SVM were limited to only dye and time duration for optimization under both GA and PSO algorithms. The selected features for each ML model and optimization algorithm are also tabulated in **Table 1** for quick reference. Moreover, the condition ranges of hyperparameters tuned via optimization techniques are detailed in **Table 2** for

recreating the functional models, while specific coding instructions in MATLAB are provided in the supplemental file.

Table 1. Feature selection for tuning ML models via GA and PSO optimization techniques.

Prediction Model	GA selected features	PSO selected features
DT	Dye, bandgap, photocatalyst dosage, solution volume, surface area, and time duration	Band gap, dye conc., photocatalyst dosage, surface area, and time duration
ELT	Dye, bandgap, surface area, and time duration	Dye, bandgap, dye conc., photocatalyst dosage, solution volume, surface area, and time duration
GPR	Dye, bandgap, dye conc., photocatalyst dosage, and time duration	Dye, dye conc., photocatalyst dosage, solution volume, surface area
SVM	Dye and time duration	Dye and time duration

In **Table 2**, a comparison of the tuned hyperparameters of machine learning models (DT, ELT, GPR, and SVM) optimized using (GA) and (PSO) are shown. For DT, GA selects a surrogate "On" and a smaller minimum leaf size (6), while PSO explores all surrogate options with a larger leaf size (12). In ELT, GA emphasizes simplicity with fewer learning cycles (10) and a higher learning rate (1), whereas PSO opts for more cycles (442.3) and a moderate learning rate (0.3826). For GPR, both methods agree on kernel parameters and functions, but GA permits higher noise (sigma = 0.0252), while PSO minimizes it (sigma = 0.0001). In SVM, GA prefers simpler models with a linear kernel, higher error tolerance, and lower kernel scale, while PSO favors a more complex polynomial kernel, stricter error control, and finer feature space granularity. Overall, GA focuses on simpler, faster configurations suitable for interpretability, while PSO seeks more precise, complex models at the cost of computational intensity.

Table 2. Tuned hyperparameters' selection of GA and PSO techniques for ML models.

	Ranges Hyperparameter	***GA***	***PSO***
DT	Surrogate [on, off]	On	all
	Minimum Leaf Size [1-152]	6	12
ELT	Number of learning cycles [10-500]	10	442.3097
	Learning rate [0.001-1]	1	0.3826

	Methods [LSBoost, Bag]	LSBoost	LSBoost
GPR	Kernel Function [Squared exponential, Ard exponential]	Ard exponential	Ard exponential
	Basic Function [Constant, Zero, Linear]	Zero	Zero
	Kernel Parameters	[1.017314230722229 3.89396463096418]	[1.01731423072222 9 3.89396463096418]
	Sigma [0.0001-0.12842]	0.0252	0.0001
SVM	Box Constraint [0.001-1000]	156.3014	41.4986
	Epsilon [1.1986e-05-1.1986]	0.0056	0.000011986
	Kernel Scale [0.001-1000]	752.4082	924.2289
	Kernel Function [Polynomial, gaussian, linear]	Linear	Polynomial

3.3 ML models performance under GA and PSO optimization

After hyperparameter tuning and feature selection, all ML models were optimized in which data was bifurcated into two sets, i.e., training and testing data. The training data points were randomly selected, comprising 80% of all data, while the remaining 20% of the data was reserved for testing the predicted model capacity. A tabulated form of data obtained for all models tested for prediction of photocatalytic rate constant before and after optimization is given in **Table 3**, while **Figure 4** depicts actual vs. predicted photocatalytic rate constants after model optimization in the form of visualized regression plots.

Table 3. Optimization parameters of ML models before and after using GA and PSO techniques.

Before Optimization		
Model	R^2	**RMSE**
DT	0.23	0.011385
ELT	0.26	0.011106
GPR	0.39	0.010151
SVM	0.36	0.010385
After optimization		
GA	**Training**	**Testing**

models				
	R^2	RMSE	R^2	RMSE
DT	0.5856	0.0010	0.4643	$1.3437e^{-17}$
ELT	0.8219	$9.5976e^{-04}$	0.9631	$5.2629e^{-18}$
GPR	0.5697	$7.3593e^{-04}$	0.2109	0.0011
SVM	0.1403	0.0213	0.0410	0.0132
PSO models	**Training**		**Testing**	
	R^2	RMSE	R^2	RMSE
DT	0.4628	$4.2423e^{-17}$	0.3382	$3.3593e^{-18}$
ELT	0.9848	0.0022	0.9921	$2.6408e^{-04}$
GPR	0.2614	$5.7188e^{-04}$	0.2287	$8.9867e^{-04}$
SVM	0.1795	0.0218	0.1795	0.0440

From the obtained data, poor prediction between experimental and predicted data was observed for all ML models before optimization as per R^2 and RSME values. Despite GPR and SVM models performing better than other models as per R^2 values, no significant correlation among the reported data could be inferred. The data was optimized according to each ML model using the GA approach, and a visual depiction of training and testing data obtained in the form of linear regression plots is given in **Figure S2**. From the figure, it can be seen that DT and GPR models have highly scattered data while SVM has ambiguous regression plots in the form of solely horizontal plotted points. In the case of the ELT model, the training dataset appears to be in a less scattered form and exhibits a more linearly correlative relationship between the x and y-axis. The R^2 values for SVM, GPR, DT, and ELT models were 0.1403, 0.5697, 0.5856, and 0.8219, respectively. The obtained prediction models were subsequently evaluated using 20% of the reserved data points. In this case, all models used input parameters obtained by the reported experimental studies in the literature and utilized each model's prediction to evaluate associated actual vs. predicted photocatalytic rate constants. It can also be seen that the SVM, GPR, and DT models' prediction capacity is insufficient and provides an erroneous prediction of photocatalytic rate constants. However, in the case of the ELT model, a statistically significant prediction of the photocatalytic rate constant is obtained, providing a highly linear correlation between the actual and predicted photocatalytic constant. The R^2 values for SVM, GPR, DT, and ELT models obtained after using the testing dataset were 0.0410, 0.2109, 0.4643, and 0.9631, respectively. The results show that the ELT model exhibited high prediction performance and can be utilized

for conducting prediction validation according to actual lab results obtained from a reference photocatalyst.

Besides GA, all ML models were also optimized using the PSO optimization method inspired by the social behavior of bird flocks. A visualized series of regression plots of all models obtained from the training and testing dataset are given in **Figure 4**, while obtained statistical metrics are given in **Table 3**. In the case of PSO optimization, DT, SVM, and GPR models tend to have more horizontally scattered plots, with high R^2 values similar to what was observed in the case of GA optimization. However, in the case of the PSO-optimized ELT model, a highly linear regressive plot was obtained from the training dataset, with a higher R^2 value (0.9846) than the GA-ELT model. Similarly, data reserved for model testing was utilized to establish the prediction capability of individual models. From **Figure 4** graphs, it is observed that the prediction of the photocatalytic rate constant is not accurate as per highly scattered data plots and high R^2 values of 0.2287, 0.1795, and 0.3382 for the models SVM, GPR, and DT, respectively. It should be noted that all models optimized using GA and PSO algorithms performed in a similar order with regard to their prediction accuracy, i.e., in descending order of ELT, DT, GPR, and SVM. Similarly, the trained predicted models were further tested against the reserved 20% testing data set to assess the model's prediction accuracy, for which a similar previous trend was observed. However, in the case of the ELT-PSO model, the highest R^2 (0.9921) value was obtained, suggesting a highly accurate prediction of the photocatalytic rate constant based on the influence of selected parameters. Considering that ELT-PSO performed better with high prediction accuracy than ELT-GA, this model was therefore selected for further experimental validation.

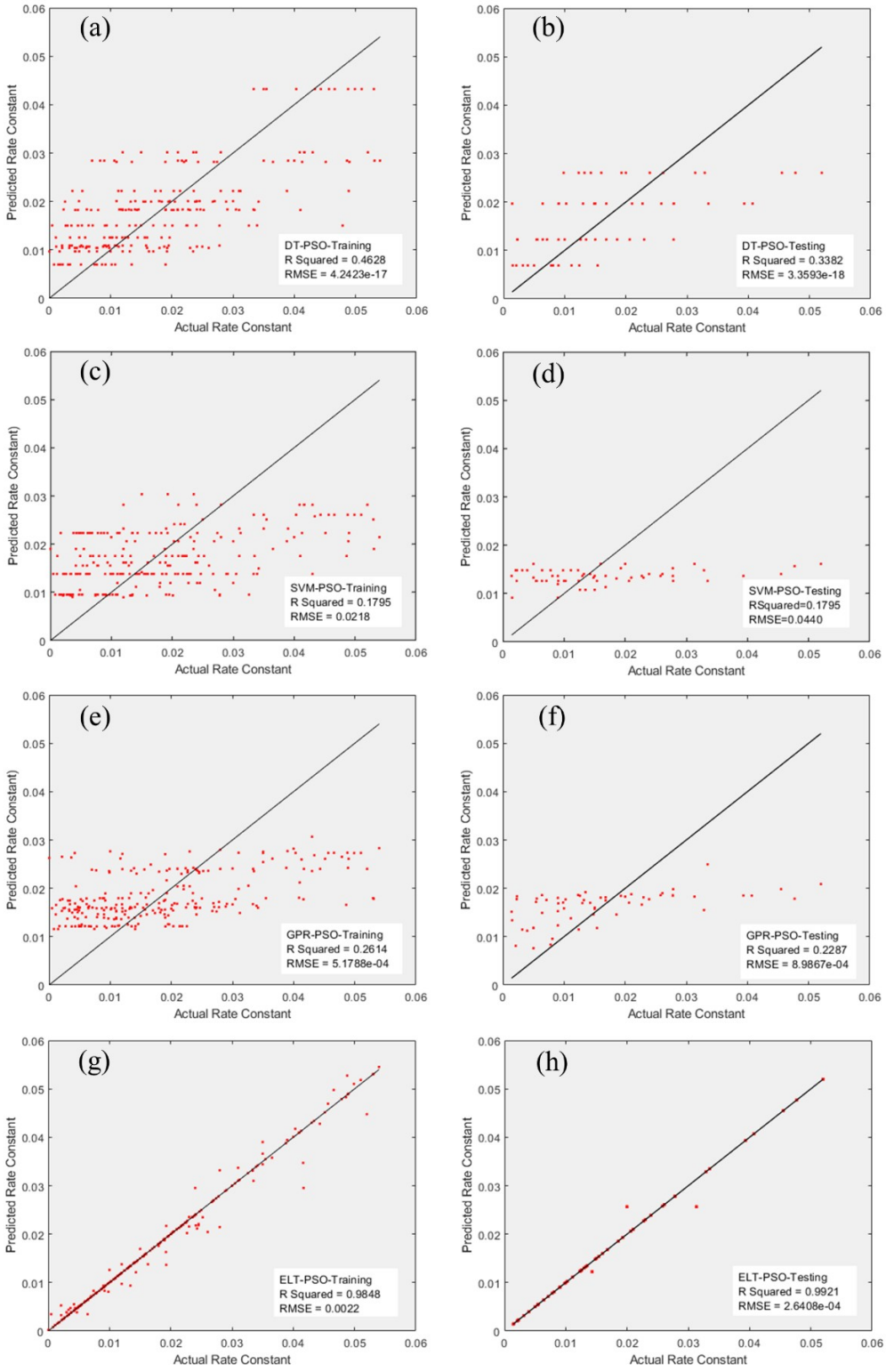

Figure 4. Illustrated plots of predicted rate constant vs. actual rate constant of cationic dyes via photocatalytic degradation using PSO technique for training (left panel) and testing (right panel) data on models DT (a-b), SVM (c-d), GPR (e-f), and ELT (g-h).

3.4 2D Partial Dependence Plots

The results of input parameters influencing the rate constant of a photocatalytic degradation process are represented in **Figure 5**. As can be seen in **Figure 5a**, the bandgap for all materials varies between 1.02-3.9, and the rate constant generally increases from 0.012 to a maximum of 0.026 value at a value around 2.9. This is due to the fundamental workings of a photocatalyst, i.e., a highly efficient photocatalyst should possess a high production of redox under visible light irradiation (Fattah-alhosseini et al., 2024). Thus, the rate of constant drops slightly from 0.026 to 0.016. **Figure 5b** elucidates the effect of photocatalyst dosage on the rate constant. Generally, the rate constant increases with an increase in the dosage of photocatalyst until an appropriate amount where saturation is achieved and the highest rate constant (0.027) is observed at an optimum value of dosage (30 mg in this case) and after this value with further increasing dosage to 50 mg, the rate constant drops to 0.020 that could be due to limited interaction of pollutants with the photocatalyst due to either agglomeration of nanoparticles leading to low surface area and less available active sites for photoactivity cause excess of nanoparticles staying without interaction (Singh et al., 2023). Similarly, the influence of surface area on rate constant is elaborated in **Figure 5c**. It is predetermined that high surface area materials possess more photodegradation sites to improve the interaction between pollutant and photocatalyst, but it also depends on the light intensity, type of light source (either UV or visible), and solution concentration of pollutants (Amano et al., 2010; He et al., 2022). Therefore, the rate constant values displayed are on the optimum values of surface areas reporting different photocatalytic materials in studies, so there is a slight fluctuation in rate constant values. **Figure 5d** illustrates the trend of rate constant vs. solution volume. A higher rate constant is observed for solutions with low volume and vice-versa when all the other parameters are kept unchanged. Moreover, an increasing and decreasing trend of rate constant is seen with an increase in solution volume from 10 to 100 mL. This is because it also depends on the concentration of pollutants in the solution. If a low solution volume contains a pollutant with a higher concentration, then a low-rate constant will be observed because photoactivity will be low (Gnanaprakasam et al., 2015). Meanwhile, at a high solution volume with a low concentration of pollutants, the rate constant would be higher due to stronger activity by the photocatalyst. Lastly, the plot in **Figure 5e** demonstrates the rate constant over a range of experimental time duration. It shows that the rate constant decreases from 0.036 to 0.015 in a nonlinear trend over a time duration of 240 min. This

is because the photocatalytic activity is highest at the start of the experiment, which tends to drop as the experiment gradually progresses until saturation is achieved or pollutants become limited in the reaction chamber after degradation. The effect on the rate constant is influenced by the simultaneous interaction of multiple input parameters. Therefore, a more detailed analysis of the rate constant is discussed in 3d plots displayed in **Figure 6**.

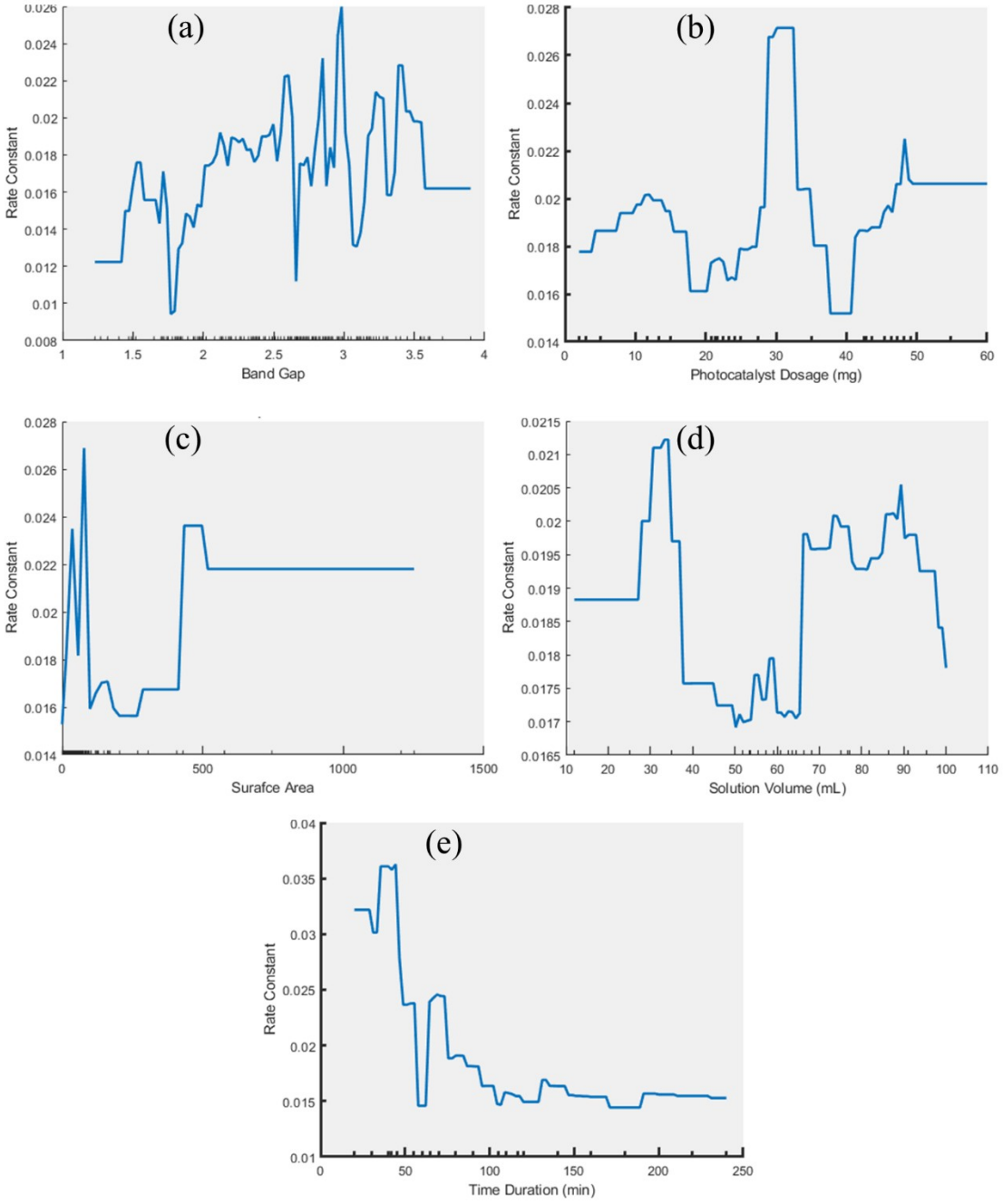

Figure 5. 2D PDPs of output rate constant vs. input experimental condition: Bandgap (a), Photocatalyst dosage (b), Surface area (c), Solution volume (d), and Time duration (e).

The effect of the interaction of dye concentration and photocatalyst dosage on the rate constant is presented in **Figure 6a.** As can be seen, the rate constant increases up to a value of ~0.08, which is obtained for dye concentration between 40 to 100 mg/L. however, for the same amount of photocatalyst dosage (below 100 mg), the rate constant rapidly rises to around 0.12 when the dye

concentration is 20 mg/L (optimum value). This could be due to more interaction of a sufficient amount of photocatalyst with the pollutant dye in the solution.

The influencing interaction of solution volume and dye concentration on the rate constant is presented in **Figure 6b.** As can be seen, the rate constant is more dependent on dye concentration rather than solution volume (yellow region). For solution volume between 200 to 1000 mL, the rate constant is about 0.08 if the dye concentration is kept low below 5 mg/L. While the rate constant is nearly 0.06-0.07 for solution volume below 50 mL, even at higher values of dye concentration. However, it is evident that the rate constant is highest at low dye concentration (< 10 mg/L) and low volume solution (< 100 mL), which is an optimum condition, giving an opportunity for the photocatalyst to degrade pollutants in solution at the highest rate (> 0.1) (Gusain et al., 2020b).

The effect on the rate constant by the interaction of dye concentration and time duration is presented in **Figure 6c.** As can be seen, when the dye concentration is increased, the rate constant is low and remains constant for all increased values of dye concentration. This is because the amount of photocatalyst is fixed and requires a longer time duration for dye degradation removal. While at dye concentration below 10 mg/L, the rate constant is 7 times high (~0.35) for an initial 200 min, represented by a spike, and this leads to a lesser time duration required to degrade and remove pollutants completely.

The influence on the rate constant by the interaction of photocatalyst dosage and solution volume is presented in **Figure 6d.** The flat region represents that no experiments were conducted on those conditions. Most of the studies were carried out for volumes up to 250 mL for all values of photocatalyst dosage (up to 1000 mg). However, experiments are concentrated at a solution volume of 100 mL and below and dosage of 100 mg and below due to batch studies with optimum degradation rate constant found to be 0.1-0.12.

The parameters of photocatalyst dosage and time duration affecting the rate constant are presented in **Figure 6e.** As can be seen, the rate constant is high at the start of the experiment, and it drops after 100 min to the remaining time of the experiment. This is due to more interaction between photocatalysts and pollutant dyes with active sites for degradation. However, that rate drops as the experiment progresses with less available active degradation sites. It is also observed that the optimum dosage of photocatalyst found in the studies is below 100 mg,

resulting in a rate constant of over 0.2 in the first hour of the interaction with the dyes, while the overall rate stays around 0.12 in the latter phase of the experiment.

The interaction effect of surface area and solution volume on the rate constant is presented in **Figure 6f.** As can be seen, solution volumes of 250 mL and below are mainly used in photodegradation studies with low optimum dye concentrations (< 30 mg/L). Therefore, it is observed that low surface areas (< 100 m^2/g) are sufficient for the photocatalytic degradation of dyes, and high surface areas (200-1000 m^2/g) do not have a significant effect on improving the rate constant (Naniwa et al., 2023).

The influence of time duration and solution volume on the rate constant is presented in **Figure 6g.** As can be seen, the highest rate constant is noticed around 0.17 in the first 100 min of the degradation removal experiments for solution volumes below 100 mL due to maximum interaction of pollutant and photocatalyst. Also, a slight drop of rate constant to 0.15 is observed for the same time duration over increased volumes of solution. A plausible reason could be that the concentration of pollutants is fixed in a solution, so the change in solution volume has less impact on the variation in the rate constant for that particular time period. Moreover, at higher solution volumes, a higher duration of time has less rate, which is because most of the interaction occurs at the start of the experiment, leading to more degradation removal and a decrease in rate constant value for the latter time due to lesser interaction (Vinothkumar et al., 2019).

The interaction of time duration and surface area influencing the rate constant is presented in **Figure 6h.** As can be seen, for all values of surface area, the rate constant has a similar trend for any particular time duration. This indicates that surface area has a lesser influence on rate constant compared to time duration. However, the rate constant is higher for low surface area photocatalysts, requiring a shorter time duration for the experiment. This is also evident in Shapley's analysis, which suggests that the rate constant does not have a direct relationship with the rate constant, and the sensitivity factor is also low, with a value of 0.5 compared to 3.5 for time duration. Furthermore, the 3D interaction plots of less influencing parameters (type of dye, band gap, and surface area) for rate constant are demonstrated in **Figure S3 and S4**.

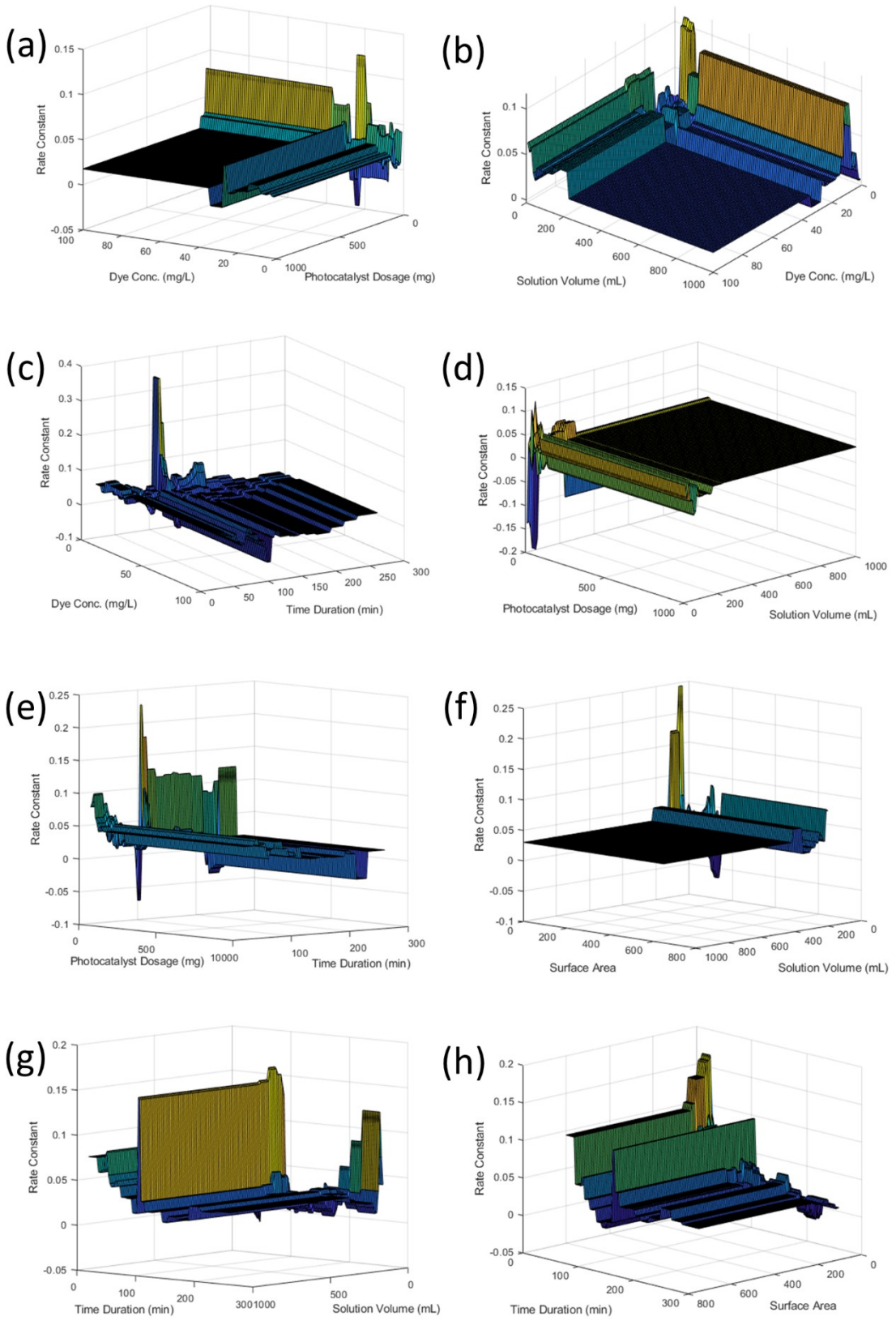

Figure 6. 3D PDPs of rate constant vs. interaction of input experimental parameters.

3.5 Shapley and Sensitivity Analysis

Shapley analysis in the form of Shapley values is widely used in ML to explain the contribution of individual parameters and their associated predictions made by a model. The concept is named after Lloyd Shapley, who introduced the idea in the 1950s. Shapley values provide a way to fairly distribute the "payout" among all features based on their relative contributions according to the concept of game theory. The Shapley analysis allows the interpretation of relevant features of the model by evaluating their equivalence using mathematically proven theory and properties about its mechanism (Roth and Bajorath, 2024). Intuitively, Shapley values of each feature, such as *j*, are obtained by iterating values of all features except *j* and then calculating the difference between the model's prediction with and without feature *j*. In essence, the Shapley value is the average contribution of a feature among all possible combinations (M. Li et al., 2024).

The SHAP analysis illustrates the dependence of input photocatalytic experimental parameters on the degradation rate constant, as shown in **Figure 7a**. Each given parameter (y-axis) has either a direct (represented by positive value) or an indirect (represented by negative value) influence on the rate constant, which is denoted by a Shapley value (x-axis). As can be seen, the time duration of photocatalytic degradation, the concentration of pollutant dye, the type of dye (cationic), and the band gap have a direct influence on the degradation rate constant in descending order. The increase in two experimental parameters, time duration and concentration of dye, allows a high number of interactions between pollutant dye and photocatalyst for more time with a greater number of active degradation sites, leading to an increase in the degradation rate constant, which results in high pollutant removal. All the other parameters (surface area, photocatalyst dosage, and solution volume) have an inverse relationship with the rate constant. The findings also corroborate with the experimental studies on the influence of those parameters. In the case of bandgap, a slight but positive correlation is observed on the rate constant. This inference is as expected since high band gap materials possess appropriate band edge positions leading to greater production of photoexcited charger carriers that increase the photodegradation rate and consequently rate constant value (Ali et al., 2023). In contrast, materials possessing low bandgap lack strong redox reaction capability, due to which the dye degradation rate constant is consequently decreased. In the case of increasing dye concentration, generally, the rate constant is increased due to strong interaction between reactive species and pollutants, after which it is reduced due to oversaturation. For instance, the photocatalytic rate constant for the degradation

of RhB dye via g-C_3N_4 was reported to increase with increasing concentration; however, at higher dye concentrations, the degradation rate constant was decreased (Vinodkumar et al., 2019). Therefore, a moderately higher dye concentration is suitable for optimal reaction rate. The same observation has been reported by another study in which the yellow dye degradation rate constant achieved via ZnO nanorods exhibited the same trend (Vidya et al., 2020). In the case of increasing photocatalyst dosage, an inverse relationship with rate constant was observed, which is in accordance with the existing studies as at higher dosage of photocatalyst, the light penetration is decreased with increased solution turbidity resulting from the screening effect (Ahmed, 2018). Summarily, the chosen features exhibited varied extent of influence on the model's prediction. This variation stems from differences in their actual influence on the photocatalytic degradation process and their statistical relevance within the sourced dataset. Features such as bandgap, photocatalyst dosage, and surface area are directly tied to the underlying photocatalytic mechanisms, such as light absorption efficiency, active site availability, and reaction kinetics, while other features such as time duration, solution volume, and dye concentrations are controlled in an experimental run resulting in varied impact. Therefore, the selected features revealed a varied impact on the ML prediction as being a combination of its physical relevance and the statistical patterns captured during training.

Sensitivity analysis is a vital tool in determining the predictive feasibility of the chosen optimal model (ELT). A leave-one-out (LOO) cross-validation was utilized to assess the performance of the ELT predictive model (Meng et al., 2024). Similar to previous models, in the ELT model's validation, the data is split into sets, i.e., testing and training. After each iteration, one data point is left out as the test set, and the remaining N-1 data points are used to train the model. The modified dataset is used to train the model and obtain predictions from the ELT model. The predictions from the ELT model obtained from containing all original datasets are compared with modified ELT models, and the error of prediction is calculated. The difference in errors in the prediction of models is then used to determine the strength of each feature. The sensitivity of each feature was assessed using LOO and is visualized in **Figure 7b** in the form of bars. This feature analysis provides useful insights into the ELT predictive model by considering the relative impact of each feature in ELT's model's prediction. This analysis aids in understanding the impact of individual features on the predictive performance of the ELT model, providing insights into feature importance.

In this context of the ELT-PSO model for predicting degradation rate constant, the feature sensitivity of time duration is highest (3.5 x 10^{-5}), followed by solution volume (1.3 x 10^{-5}), and then the remaining features (surface area, photocatalyst dosage, and band gap) below 0.5 x 10^{-5}. On the other hand, the type of dye and concentration of dye are not sensitive to the rate constant prediction by the model, which indicates that prediction would not significantly change by varying these parameters. The results suggest that slight variations in the time duration of photocatalytic activity can significantly impact the outcome of the prediction. At the same time, the change in the band gap has minimal effect on the prediction values of the rate constant. Thus, the results provide a favorable understanding of the interpretations required to prioritize the focus on parameters critical for designing an optimized photocatalytic degradation experiment.

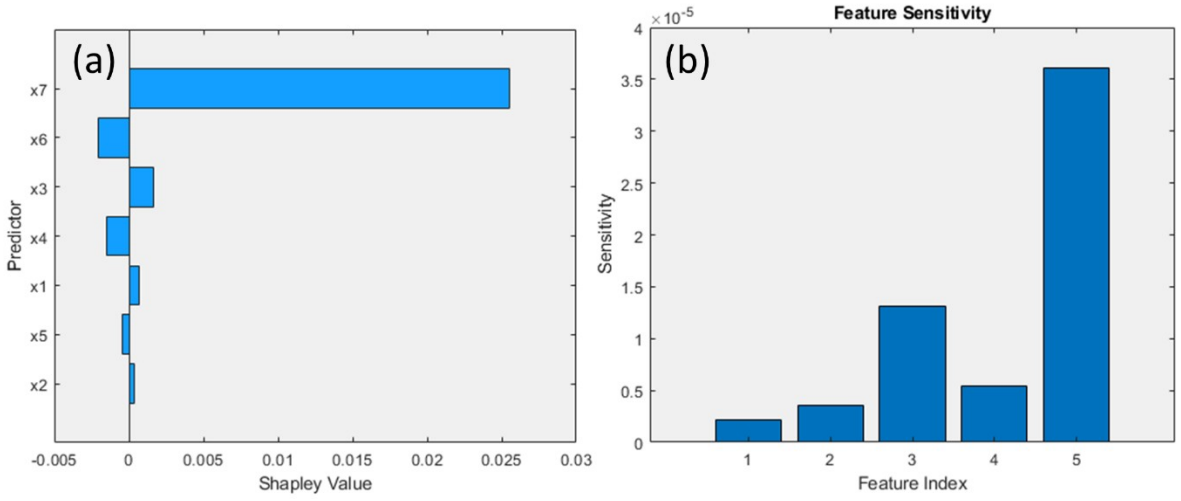

Figure 7. Shapley plot (x1: type of dye, x2: bandgap, x3: dye concentration, x4: dosage, x5: solution volume, x6: surface area, and x7: time duration) (a) and Feature sensitivity plot (1: bandgap, 2: dosage, 3: solution volume, 4: surface area, and 5: time duration) (b).

3.6 Experimental validation

To establish actual model prediction and to validate the performance of obtained trained ELT-PSO, which provided high prediction according to performance metrics (R^2 and RMSE), we utilized the previously published experimental data from the results obtained in our lab on photodegradation of MB using Cu_2O(x wt. %)/WO_3 heterostructure photocatalysts as reference material, with detailed data available in our previous report (Ali et al., 2021). The actual obtained results for the sample are displayed in **Table 4**. From the table, it can be seen that the predicted rate constant and experimentally obtained rate constant closely lie within a low error margin, indicating a high model's performance and accuracy. The experimental data were obtained with

a sole variation of bandgaps (2.62-2.8) owing to the different synthesized variations of Cu_2O/WO_3. The obtained model thus can provide a generalized prediction of rate constants based on selected essential experimental features.

Table 4. Developed model's performance evaluation for validation by laboratory experimental results for MB degradation/discoloration.

Band Gap (eV)	Dye Conc. (mg/L)	Photocatalyst dosage (mg)	Solution volume (mL)	Specific surface area ($m^2.g^{-1}$)	Time duration (min)	Actual rate constant (min^{-1})	Predicted rate constant	Error %
2.62	3.5	5	12	9.1	240	0.009	0.008676	3.6
2.68	3.5	10	12	10.7	240	0.0021	0.0020307	3.3
2.76	3.5	20	12	11.3	240	0.0024	0.0024972	4.05
2.8	7.0	5	12	14.3	240	0.0031	0.0031899	2.9
3.25	3.7	5	12	10.2	75	0.02129	0.02207773	3.7
3.25	3.7	10	12	10.2	75	0.03469	0.03610535	4.08
3.25	7.0	5	12	10.2	150	0.01207	0.01165962	3.4
3.25	7.0	10	12	10.2	150	0.0175	0.0181825	3.9

To further substantiate the model's validity, pristine ZnO as a model photocatalyst was prepared according to the methodology described in section 1.1, while the structural and optical characterization are given in **Figure 8**. The as-prepared sample has been thoroughly characterized in the previous report therefore only general characterization is deemed necessary. From **Figure 8a**, in a zoomed-out look, tiny ZnO blocks are assembled into large blocks, which also appear to be consistent in formation throughout, while a closer look in **Figure 8b** reveals the well-defined morphology of the obtained sample, which appears to be tiny blocks uniformly distributed throughout the view. From **Figure 8c**, all XRD characteristic peaks associated with ZnO can be seen, including the most intense peaks at 2θ 37.10, 40.22, 42.20 corresponding to the (100), (002), and (101) ZnO hexagonal structure (Gull et al., 2024). No presence of any impurity phases was observed in the XRD diffractogram, indicating structural purity. Extended XRD peaks survey, including FWHM values and crystallite size calculated according to the Halder-Wagner method (Nath et al., 2020), are provided in **Table S3**. Besides that, XRD and SEM, the optical bandgap of ZnO was evaluated using the Tauc plot as shown in **Figure 8c**. By extrapolating the graph to 0, the bandgap at the *x*-axis was obtained to be ~ 3.25 eV. The as-prepared sample was used to validate the model at selected random variable conditions, and the

obtained results are given in **Table 4**. The as-prepared ZnO sample was subsequently used to validate the ELT-PSO model under varied conditions of dye concentration (3.7 and 7 mg/L) and photocatalyst dosage (5 and 10 mg). The obtained constant rate prediction values were in close proximity to the experimentally obtained values with an error margin of less than 4%, thus confirming the suitability of the developed model. This error can be reasonably associated with the exclusion of other experimental variables, such as light intensity and distance from the irradiation source, which are usually omitted by literature studies. The developed hybrid ELT-PSO model provides useful insights into experimental features highly influencing the photocatalytic dye degradation process, as well as a highly accurate prediction of the photocatalytic rate constant for the removal of dye via photocatalysis. Summarily, the current study complements and advances the findings of utilizing ML approach for photocatalysis aspects. To provide a general overview of the current scope of our study in comparison to existing literature reports pertaining to the ML-based photocatalytic prediction for dye removal, **Table S4** is provided in the supplemental section. From the table, it can be seen that each published study provides a different ML approach, models, and sourced dataset. Considering the existing reports, our study provides additional useful insights for the prediction of photocatalytic rate constant by variation in selected experimental parameters of various types of reported photocatalysts. Moreover, in contrast to most other reports, our experimental validation on the dataset, which was not included in the original sourced dataset, improves the reliability of the developed ML model.

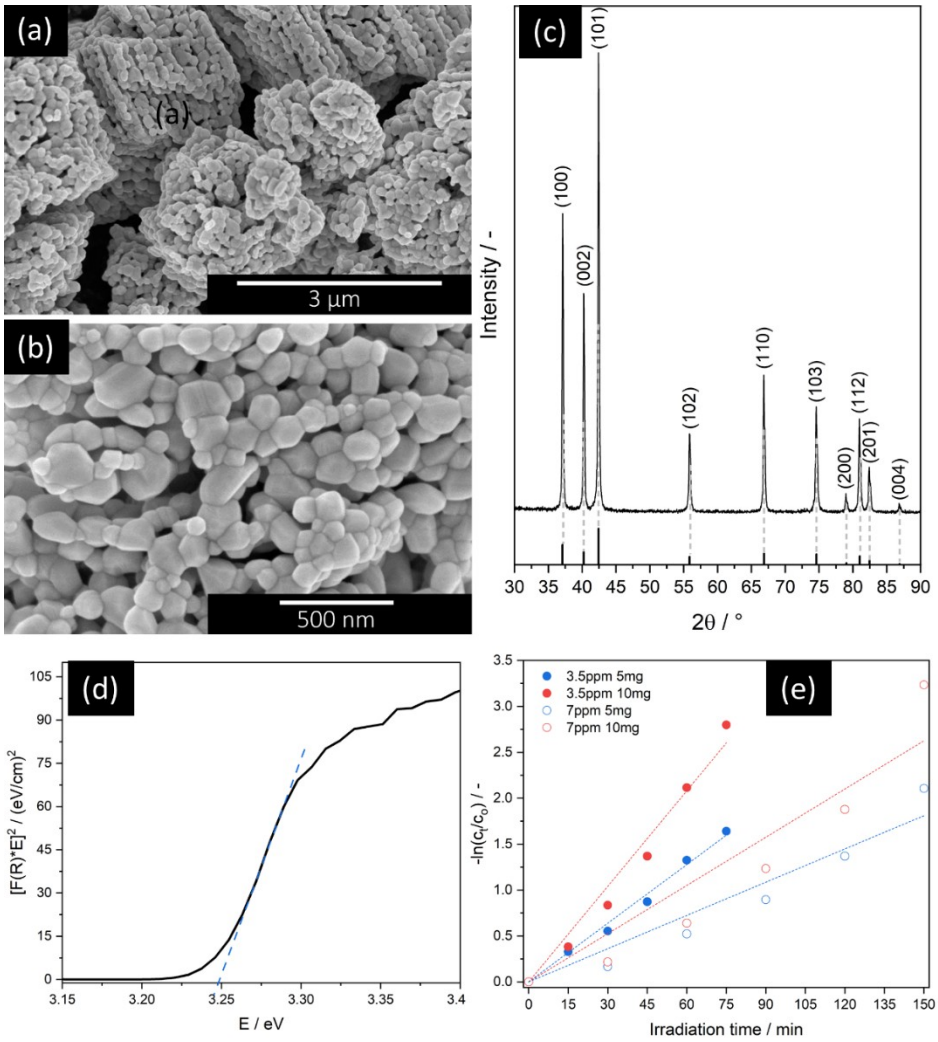

Figure 8. SEM characterization (a-b), XRD (c), Tauc plot (d), and photocatalytic degradation results (e) of the as-prepared ZnO sample utilized for experimental validation of the ML model.

4. Conclusions

The photocatalytic degradation removal of cationic dyes was successfully modeled using an ML strategy to predict the degradation rate constant. It was found that DT, ELT, GPR, and SVM had R^2 values of 0.23, 0.26, 0.39, and 0.36 before optimization, respectively. The models were optimized using GA and PSO techniques to fine-tune hyperparameters, and after feature selection, the value of R^2 was improved for all models and showed relatively better values with PSO. Although feature selection differed using GA and PSO techniques, ELT-aligned PSO outperformed all the models for both training (R^2 = 0.9848, and RMSE = 0.0022) and testing data (R^2 = 0.9921, and RMSE = 2.6408e^{-04}). The partial dependence plots elucidated that solution volume, dye concentration, photocatalyst dosage, and time duration are essential parameters for

determining rate constant, while Shapley's analysis further revealed time duration and dye concentration to be the most important parameters along with solution volume and time duration of the experiment to be most sensitive parameters affecting prediction of rate constant for photocatalytic degradation reaction. Furthermore, additional validation experiments on different synthesized heterostructure photocatalyst nanoparticles demonstrated precise results with a 4.05% maximum error difference between experimental and predicted rate constant values. Summarily, this study offers useful insights for efficient and effective photocatalytic removal of cationic dyes from wastewater using an ML-based approach.

Acknowledgments

This work was supported by the European Just Transition Fund within the Operational Programme: Just Transition under the aegis of the Ministry of the Environment of the Czech Republic, project CirkArena number CZ.10.03.01/00/22_003/0000045 and Operational Programme Johannes Amos Comenius OP JAC "Application potential development in the field of polymer materials in the context of circular economy compliance (POCEK)", number CZ.02.01.01/00/23_021/0009004. The authors further acknowledge the financial support from the Ministry of Education, Youth, and Sports of the Czech Republic - DKRVO (RP/CPS/2024-28/002 and RP/CPS/2024-28/007). Hassan Ali, Muhammad Yasir, and Ali Can Guler also express their gratitude for support within the "Creativity, Intelligence & Talent for the Zlin Region" (CIT - ZK) program.

Conflict of interest

The authors declare there exists no form of competing interests.

Data availability

The data that support the findings of this study are openly available in Zenodo at

https://doi.org/10.5281/zenodo.13843658

References

Ahmed, L.M., 2018. Photo-Decolourization Kinetics of Acid Red 87 Dye in ZnO Suspension Under Different Types of UV-A Light. Asian J. Chem. 30, 2134–2140. https://doi.org/10.14233/ajchem.2018.21520

Akram, M.Y., Ashraf, T., Tong, L., Yin, X., Dong, H., Lu, H., 2024. Architecting high-performance photocatalysts: A review of modified 2D/2D graphene/g-C3N4 heterostructures. Journal of Environmental Chemical Engineering 12, 113415. https://doi.org/10.1016/j.jece.2024.113415

Akshhayya, C., Ansari, M.O., Janani, B., Subhiksha, V., Melaibari, A.A., Hasan, P.M.Z., Asad, M., Sudheer Khan, S., 2023. Construction of a S-scheme CdS/$CuFe_2O_4$ heterojunction with suppressed charge recombination for enhanced visible light-driven photocatalytic activity. New J. Chem. 47, 5431–5440. https://doi.org/10.1039/D2NJ05616J

Akyildiz, H.I., Yigit, E., Arat, A.B., Islam, S., 2024. A machine learning approach for the estimation of photocatalytic activity of ALD ZnO thin films on fabric substrates. Journal of Photochemistry and Photobiology A: Chemistry 448, 115308. https://doi.org/10.1016/j.jphotochem.2023.115308

Ali, H., Guler, A., Masar, M., Urbanek, P., Urbanek, M., Skoda, D., Suly, P., Machovsky, M., Galusek, D., Kuritka, I., 2021. Solid-State Synthesis of Direct Z-Scheme Cu2O/WO3 Nanocomposites with Enhanced Visible-Light Photocatalytic Performance. Catalysts 11, 293. https://doi.org/10.3390/catal11020293

Ali, H., Masar, M., Yasir, M., Machovsky, M., Monteiro, O.C., Kuritka, I., 2023. Current trends in environmental and energy photocatalysis and ISO standardization. Journal of Environmental Chemical Engineering 11, 111541. https://doi.org/10.1016/j.jece.2023.111541

Amano, F., Nogami, K., Tanaka, M., Ohtani, B., 2010. Correlation between Surface Area and Photocatalytic Activity for Acetaldehyde Decomposition over Bismuth Tungstate Particles with a Hierarchical Structure. Langmuir 26, 7174–7180. https://doi.org/10.1021/la904274c

Anjaneyulu, B., Goswami, S., Banik, P., Chauhan, V., Raghav, N., Chinmay, 2024. Revolution of Artificial Intelligence in Computational Chemistry Breakthroughs. Chemistry Africa. https://doi.org/10.1007/s42250-024-00976-5

Bakır, R., Orak, C., Yüksel, A., 2024. A machine learning ensemble approach for predicting solar-sensitive hybrid photocatalysts on hydrogen evolution. Phys. Scr. 99, 076015. https://doi.org/10.1088/1402-4896/ad562a

Bang Truong, H., Cuong Nguyen, X., Hur, J., 2023. Recent advances in g–C3N4–based photocatalysis for water treatment: Magnetic and floating photocatalysts, and applications of machine-learning techniques. Journal of Environmental Management 345, 118895. https://doi.org/10.1016/j.jenvman.2023.118895

Banisheikholeslami, A., Qaderi, F., 2024. Applied machine learning to the determination of biochar hydrogen sulfide adsorption capacity. Mach Learn 113, 3419–3441. https://doi.org/10.1007/s10994-023-06446-2

Baskar, G., Nashath Omer, S., Saravanan, P., Rajeshkannan, R., Saravanan, V., Rajasimman, M., Shanmugam, V., 2024. Status and future trends in wastewater management strategies using artificial intelligence and machine learning techniques. Chemosphere 362, 142477. https://doi.org/10.1016/j.chemosphere.2024.142477

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324
Chow, V., Phan, R.C.-W., Ngo, A.C.L., Krishnasamy, G., Chai, S.-P., 2022. Data-driven photocatalytic degradation activity prediction with Gaussian process. Process Safety and Environmental Protection 161, 848–859. https://doi.org/10.1016/j.psep.2022.03.020
Fattah-alhosseini, A., Sangarimotlagh, Z., Karbasi, M., 2024. Latest progress in photocatalytic hydrogen production using MXene (Ti3C2)/MOFs composite: A review. International Journal of Hydrogen Energy 79, 771–790. https://doi.org/10.1016/j.ijhydene.2024.07.058
Freiesleben, T., König, G., Molnar, C., Tejero-Cantero, Á., 2024. Scientific Inference with Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena. Minds & Machines 34, 32. https://doi.org/10.1007/s11023-024-09691-z
Gao, Z., Ren, Z., Cui, T., Fu, Y., 2025. Machine learning-based analysis of microplastic-induced changes in anaerobic digestion parameters influencing methane yield. Journal of Environmental Management 377, 124627. https://doi.org/10.1016/j.jenvman.2025.124627
Gnanaprakasam, A., Sivakumar, V.M., Thirumarimurugan, M., 2015. Influencing Parameters in the Photocatalytic Degradation of Organic Effluent via Nanometal Oxide Catalyst: A Review. Indian Journal of Materials Science 2015, 1–16. https://doi.org/10.1155/2015/601827
Gull, F., Riaz, R., Ansari, K., Atiq, H., 2024. Examining the photo catalytic potency of annealed and un-annealed ZnO and nickel doped ZnO for degradation of organic pollutants in waste waters. Sci Rep 14, 21828. https://doi.org/10.1038/s41598-024-60258-5
Gusain, R., Kumar, N., Ray, S.S., 2020a. Factors Influencing the Photocatalytic Activity of Photocatalysts in Wastewater Treatment, in: Fosso-Kankeu, E., Pandey, S., Ray, S.S. (Eds.), Photocatalysts in Advanced Oxidation Processes for Wastewater Treatment. Wiley, pp. 229–270. https://doi.org/10.1002/9781119631422.ch8
Gusain, R., Kumar, N., Ray, S.S., 2020b. Factors Influencing the Photocatalytic Activity of Photocatalysts in Wastewater Treatment, in: Fosso-Kankeu, E., Pandey, S., Ray, S.S. (Eds.), Photocatalysts in Advanced Oxidation Processes for Wastewater Treatment. Wiley, pp. 229–270. https://doi.org/10.1002/9781119631422.ch8
Hancock, J.T., Bauder, R.A., Wang, H., Khoshgoftaar, T.M., 2023. Explainable machine learning models for Medicare fraud detection. J Big Data 10, 154. https://doi.org/10.1186/s40537-023-00821-5
He, X., Yang, Y., Li, Y., Chen, J., Yang, S., Liu, R., Xu, Z., 2022. Effects of structure and surface properties on the performance of ZnO towards photocatalytic degradation of methylene blue. Applied Surface Science 599, 153898. https://doi.org/10.1016/j.apsusc.2022.153898
Hua, R., Huang, Y., Xia, Y., She, H., Wang, L., Huang, J., Wang, Q., 2024. Enhanced photocatalytic CO2 reduction over Z-scheme β-Ga2O3/TiO2 heterojunction composite catalyst: Synthesis, performance, and mechanism. Journal of Environmental Chemical Engineering 12, 113135. https://doi.org/10.1016/j.jece.2024.113135
Imai, K., Fukushima, T., Heguri, S., Dohshi, S., Takahashi, M., Higashimoto, S., 2024. Boosting visible-light response for the complete decomposition of volatile organic compounds on the Cu-oxide deposited WO3 photocatalyst by the synergistic effects of TiO2. Journal of Environmental Chemical Engineering 12, 113610. https://doi.org/10.1016/j.jece.2024.113610

Jiang, Z., Hu, J., Tong, M., Samia, A.C., Zhang, H. (Judy), Yu, X. (Bill), 2021. A Novel Machine Learning Model to Predict the Photo-Degradation Performance of Different Photocatalysts on a Variety of Water Contaminants. Catalysts 11, 1107. https://doi.org/10.3390/catal11091107

Johar, M.A., Afzal, R.A., Alazba, A.A., Manzoor, U., 2015. Photocatalysis and Bandgap Engineering Using ZnO Nanocomposites. Advances in Materials Science and Engineering 2015, 1–22. https://doi.org/10.1155/2015/934587

K C, A., Rao, C.S., Nair, V., 2024. Combination of ensemble machine learning models in photocatalytic studies using nano TiO2 - Lignin based biochar. Chemosphere 352, 141326. https://doi.org/10.1016/j.chemosphere.2024.141326

Khan, M., Ullah, Z., Mašek, O., Raza Naqvi, S., Nouman Aslam Khan, M., 2022. Artificial neural networks for the prediction of biochar yield: A comparative study of metaheuristic algorithms. Bioresource Technology 355, 127215. https://doi.org/10.1016/j.biortech.2022.127215

Kim, C.-M., Jaffari, Z.H., Abbas, A., Chowdhury, M.F., Cho, K.H., 2024. 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. https://doi.org/10.1016/j.jhazmat.2023.132995

Knott, M., Perez-Cruz, F., Defraeye, T., 2023. Facilitated machine learning for image-based fruit quality assessment. Journal of Food Engineering 345, 111401. https://doi.org/10.1016/j.jfoodeng.2022.111401

Lemm, D., Falk Von Rudorff, G., Anatole Von Lilienfeld, O., 2023. Improved decision making with similarity based machine learning: applications in chemistry. Mach. Learn.: Sci. Technol. 4, 045043. https://doi.org/10.1088/2632-2153/ad0fa3

Li, J., Chen, Z., Wu, J., Lin, J., He, P., Zhu, R., Peng, C., Zhang, H., Li, W., Fang, X., Shen, H., 2023. Constructing a link between multivariate titanium-based semiconductor band gaps and chemical formulae based on machine learning. Materials Today Communications 35, 106299. https://doi.org/10.1016/j.mtcomm.2023.106299

Li, J., Li, Y., Selishchev, D., Zhang, G., 2024. Near-infrared responsive photocatalysts for environmental remediation and energy conversion: A review. Chemosphere 367, 143599. https://doi.org/10.1016/j.chemosphere.2024.143599

Li, L., Zhang, P., Li, N., Reyila, T., Yu, Y., Su, X., Peng, C., Han, L., 2024. Application of g-C3N4-based photocatalysts for N2 photofixation. Journal of Environmental Chemical Engineering 12, 112142. https://doi.org/10.1016/j.jece.2024.112142

Li, M., Sun, H., Huang, Y., Chen, H., 2024. Shapley value: from cooperative game to explainable artificial intelligence. Auton. Intell. Syst. 4, 2. https://doi.org/10.1007/s43684-023-00060-8

Long, X., Huangfu, X., Huang, R., Liang, Y., Wu, S., Wang, J., 2024. The application of machine learning methods for prediction of heavy metal by activated carbons, biochars, and carbon nanotubes. Chemosphere 354, 141584. https://doi.org/10.1016/j.chemosphere.2024.141584

Luo, H., Yan, J., Shan, Y., Zhou, J., Yu, J., Boury, B., Wu, H., Xiao, H., Huang, L., Yuan, Z., Chen, L., 2022. Metal-free photocatalyst for nitrogen fixation under visible light based on COF/g-C3N4/ CNT nanocomposite. Journal of Environmental Chemical Engineering 10, 107713. https://doi.org/10.1016/j.jece.2022.107713

Meng, Z., Wang, J., Lin, L., Wu, C., 2024. Sensitivity analysis with iterative outlier detection for systematic reviews and meta-analyses. Statistics in Medicine 43, 1549–1563. https://doi.org/10.1002/sim.10008

Naniwa, S., Kato, K., Yamakata, A., Yamamoto, A., Yoshida, H., 2023. A Quantitative Study on the Relationship of Specific Surface Area and Carrier Lifetime to Photocatalytic Activity of Anatase TiO_2 Nanoparticles. ACS Catal. 13, 15212–15218. https://doi.org/10.1021/acscatal.3c04209

Nath, D., Singh, F., Das, R., 2020. X-ray diffraction analysis by Williamson-Hall, Halder-Wagner and size-strain plot methods of CdSe nanoparticles- a comparative study. Materials Chemistry and Physics 239, 122021. https://doi.org/10.1016/j.matchemphys.2019.122021

Ollis, D.F., 2018. Kinetics of Photocatalyzed Reactions: Five Lessons Learned. Front. Chem. 6, 378. https://doi.org/10.3389/fchem.2018.00378

Özsoysal, S., Oral, B., Yıldırım, R., 2024. Analysis of photocatalytic CO_2 reduction over MOFs using machine learning. J. Mater. Chem. A 12, 5748–5759. https://doi.org/10.1039/D3TA07001H

Quinlan, J.R., 1986. Induction of Decision Trees. Machine Learning 1, 81–106. https://doi.org/10.1023/A:1022643204877

Rashid, R., Shafiq, I., Gilani, M.R.H.S., Maaz, M., Akhter, P., Hussain, M., Jeong, K.-E., Kwon, E.E., Bae, S., Park, Y.-K., 2024. Advancements in TiO2-based photocatalysis for environmental remediation: Strategies for enhancing visible-light-driven activity. Chemosphere 349, 140703. https://doi.org/10.1016/j.chemosphere.2023.140703

Roth, J.P., Bajorath, J., 2024. Machine learning models with distinct Shapley value explanations decouple feature attribution and interpretation for chemical compound predictions. Cell Reports Physical Science 5, 102110. https://doi.org/10.1016/j.xcrp.2024.102110

Sacco, O., Vaiano, V., Sannino, D., 2020. Main parameters influencing the design of photocatalytic reactors for wastewater treatment: a mini review. J of Chemical Tech & Biotech 95, 2608–2618. https://doi.org/10.1002/jctb.6488

Saeidpour, S., Khoshnevisan, B., Boroumand, Z., 2022. Synthesis and characterization of a g-C3N4/TiO2-ZnO nanostructure for photocatalytic degradation of methylene blue. Nano Futures 6, 035001. https://doi.org/10.1088/2399-1984/ac74fa

Sedlak, J., Kuritka, I., Masar, M., Machovsky, M., Urbanek, P., Bazant, P., Janota, P., Dvorackova, M., 2020. Contributions of morphological and structural parameters at different hierarchical morphology levels to photocatalytic activity of mesoporous nanostructured ZnO. Applied Surface Science 513, 145773. https://doi.org/10.1016/j.apsusc.2020.145773

Shabna, S., Singh, C.J.C., Dhas, S.D.S.J., Jeyakumar, S.C., Biju, C.S., 2024. An overview of prominent factors influencing the photocatalytic degradation of cationic crystal violet dye employing diverse nanostructured materials. J of Chemical Tech & Biotech 99, 1027–1055. https://doi.org/10.1002/jctb.7593

Siddique, M., Khan, N.M., Saeed, M., 2019. Photocatalytic Activity of Bismuth Ferrite Nanoparticles Synthesized via Sol-Gel Route. Zeitschrift für Physikalische Chemie 233, 595–607. https://doi.org/10.1515/zpch-2018-1225

Singh, G., Ubhi, M.K., Jeet, K., Singla, C., Kaur, M., 2023. A Review on Impacting Parameters for Photocatalytic Degradation of Organic Effluents by Ferrites and Their Nanocomposites. Processes 11, 1727. https://doi.org/10.3390/pr11061727

Sun, N., Si, X., He, L., Zhang, J., Sun, Y., 2024. Strategies for enhancing the photocatalytic activity of semiconductors. International Journal of Hydrogen Energy 58, 1249–1265. https://doi.org/10.1016/j.ijhydene.2024.01.319

Taye, M.M., 2023. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. Computers 12, 91. https://doi.org/10.3390/computers12050091

Tu, Z., Stuyver, T., Coley, C.W., 2023. Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery. Chem. Sci. 14, 226–244. https://doi.org/10.1039/D2SC05089G

Verma, S., Tirumala Rao, B., Singh, R., Kaul, R., 2021. Photocatalytic degradation kinetics of cationic and anionic dyes using Au–ZnO nanorods: Role of pH for selective and simultaneous degradation of binary dye mixtures. Ceramics International 47, 34751–34764. https://doi.org/10.1016/j.ceramint.2021.09.014

Vidya, C., Manjunatha, C., Sudeep, M., Ashoka, S., Lourdu Antony Raj, M.A., 2020. Photo-assisted mineralisation of titan yellow dye using ZnO nanorods synthesised via environmental benign route. SN Appl. Sci. 2, 743. https://doi.org/10.1007/s42452-020-2537-2

Vinodkumar, T., Subramanyam, P., Kumar, K.V.A., Reddy, B.M., Subrahmanyam, C., 2019. Construction of metal oxide decorated g-C3N4 materials with enhanced photocatalytic performance under visible light irradiation. J Chem Sci 131, 13. https://doi.org/10.1007/s12039-018-1588-z

Vinothkumar, P., Manoharan, C., Shanmugapriya, B., Bououdina, M., 2019. Effect of reaction time on structural, morphological, optical and photocatalytic properties of copper oxide (CuO) nanostructures. J Mater Sci: Mater Electron 30, 6249–6262. https://doi.org/10.1007/s10854-019-00928-7

Weinreich, J., Falk Von Rudorff, G., Anatole Von Lilienfeld, O., 2023. Encrypted machine learning of molecular quantum properties. Mach. Learn.: Sci. Technol. 4, 025017. https://doi.org/10.1088/2632-2153/acc928

Xiong, H.-Q., Bao, H.-R., Long, F., Du, Y.-Y., Qu, J.-Z., Luan, Z.-X., Sun, X.-L., 2024. From lab to nature: Overcoming challenges in applying in-situ photocatalysis to water bodies. Journal of Environmental Chemical Engineering 12, 112656. https://doi.org/10.1016/j.jece.2024.112656

Zhai, X., Chen, M., 2023. A machine learning-based nano-photocatalyst module for accelerating the design of Bi_2 WO_6 /MIL-53(Al) nanocomposites with enhanced photocatalytic activity. Nanoscale Adv. 5, 4065–4073. https://doi.org/10.1039/D3NA00122A

Zheng, H., Wu, Y., 2019. A XGBoost Model with Weather Similarity Analysis and Feature Engineering for Short-Term Wind Power Forecasting. Applied Sciences 9, 3019. https://doi.org/10.3390/app9153019