A predictive model for degradation of doped TiO_2 photocatalytic wastewater pollutants via machine learning

Qing Liu[1], Kewei Pan[1], Ying Lu[1], Wei Wei[1], Shixing Wang[2], Wenjia Du[3], Zhao Ding[4], Yang Zhou[1*]

[1] School of Textile Science and Engineering / State Key Laboratory of New Textile Materials and Advanced Processing Technology, Wuhan Textile University, Wuhan 430200, China

[2] Faculty of Metallurgical and Energy Engineering, Kunming University of Science and Technology, Kunming 650093, China

[3] Electrochemical Innovation Lab, Department of Chemical Engineering, University College London, London, WC1E 7JE, UK

[4] College of Materials Science and Engineering, Chongqing University, Chongqing 400044, China

Corresponding author Email address: yzhou@wtu.edu.cn (Y. Zhou)

Abstract

TiO_2 photocatalytic degradation technology, as an efficient, clean technology, is widely used in the treatment of contaminated wastewater. To expand the absorption spectrum of TiO_2, from UV to visible light, considerable effort has been put into the modification of TiO_2. Current TiO_2 studies still rely on experimental work, mainly focusing on the effects of experimental variables on photocatalytic degradation. However, multiple

variables introduce the experimental complexity and increase the cost. As a result, TiO_2 development could be time-consuming and less cost-effective. Herein, we use a machine learning (ML) approach to study the photocatalytic degradation of doped TiO_2. In this study, the degradation rate of pollutant solution in the presence of doped TiO_2 was simulated under various experimental conditions using a LightGBM model. The training data from experiments comprised nine inputs, namely dopant, dopant/Ti molar ratio, calcination temperature, pollutant, catalyst/pollutant mass ratio, pH, experimental temperature, light wavelength and illumination time. The predicted result from trained model shows a good accuracy with a high coefficient of determination (92.8%). By ranking importance of these influencing variables, this study may help a better design of TiO_2 for wastewater treatment, thereby improving the purification efficiency and saving natural resources.

Keywords: Doped TiO_2; Photodegradation of wastewater; Machine learning; LightGBM; Degradation rate.

1. Introduction

Recently, water shortages have become an ever-growing challenge as the rapid industrial growth, environmental pollution, depleted water resources [1]. Currently, the chemical pollutant in circulation is 38,000 with more than 300 new materials being synthesized every year [2]. Consequently, it is difficult to purify the wastewater as its unstable quality with the large discharging, high chroma, and high content of refractory organics. Traditional wastewater treatment methods struggle to address above issues. For

example, the physical adsorption only separates pollutants without complete decomposition and might introduce the secondary pollution [3]. High-cost chemical methods have poor decolorization effects [4]. Hence, there is a pressing need to develop an efficient, green, and cost-effective technology for wastewater purification.

The newly developed approach via photocatalytic degradation is a promising technology to address the challenge of wastewater treatment. Of the most interest, the TiO_2, a well-known photocatalyst, has been attracted considerable attention from academics and industrial sectors. Because of its exceptional material properties, including the high refractive index and ultraviolet (UV) absorption, excellent incident photoelectric conversion efficiency and dielectric constant, good photocatalytic activity, photostability, chemical stability, and long-term corrosion resistance and nontoxicity [5]. When TiO_2 absorbs external energy under UV irradiation, the electrons (e^-) in the valence band (VB) will become excited to the conduction band (CB) and migrate to the particle surface under the electric field, resulting in the generation of the electron-hole pairs. The e^- on the TiO_2 surface could be easily captured by oxidizing substances such as dissolved oxygen in water. While the holes (h^+) in the VB could be captured by the OH^- and H_2O to generate the hydroxyl radicals (·OH). The oxidation could purify most of the organic and inorganic pollutants in wastewater and mineralize them into harmless substances (i.e., inorganic small molecules, CO_2 and H_2O). This mechanism could be illustrated by Equations (1) - (5) and Figure 1.

$$TiO_2 + hv \rightarrow h^+ + e^- \quad (1)$$

$$e^- + O_2 \rightarrow \cdot O_2^- \quad (2)$$

$$h^+ + OH^- \rightarrow \cdot OH \quad (3)$$

$$h^+ + H_2O \rightarrow \cdot OH + H^+ \quad (4)$$

$$\cdot OH + h^+ + \text{Organic} \rightarrow \text{Intermediates} \rightarrow CO_2 + H_2O \quad (5)$$

Unfortunately, Ti has a large band gap (3.2 eV); therefore, only a small fraction of solar light (approximately 5%) in the UV region can be utilized [6]. To develop a practical TiO_2-based photocatalytic process, the efforts have been focused on either improving its photocatalytic activity or extending its absorption spectrum (from the UV to the visible range) by adding a second element to the TiO_2 bulk structure [7]. Both metal and non-metal element doping TiO_2 were used to impede electron-hole recombination and thus increase both visible light and capacity.

Park et al. [8] synthesized N-TiO_2 nanostructured materials via a graft polymerization. The resultant materials have enhanced catalytic activity for the degradation of methylene orange (MO) dyes after irradiation with visible light. This is attributed to the incorporation of N into the TiO_2 structure, which reforms the electronic band level of TiO_2. The doped material could absorb visible light, but e^-/h^+ pair recombination is limited because of the strong catalyzation. M. Khairy et al. [9] prepared M-doped TiO_2 nanoparticles (M = Cu, Zn) by the sol–gel method. The photocatalytic activities for methyl orange (MO) degradation and chemical oxygen demand (COD) were investigated. An optical study

showed that doping ions led to an increase in the absorption edge wavelength and a decrease in the band gap energy of TiO_2. In general, the doped TiO_2 demonstrated higher photocatalytic activities than no doped one. It suggests that a host of reaction conditions (such as, dopant loading, initial concentration of pollutant solution and calcination temperature) could affect the degradation rate of the catalyst through the photocatalytic process [10-11]. However, it is challenging to determine the most efficient preparation conditions for the doped sample and the experimental conditions for photoreactivity assays over a wide range of variations in experimental variables. In the past, conventional research routes usually rely on the literature review and carry out a series of experiments that have not been explored (trial and error). Consequently, research progress might be restricted due to the excessive number of publications and the misleading data from different sources. Application of a ML approach to analyze the large amount of data collected from the previous literature could reveal unprecedented information, such as valuable trends.

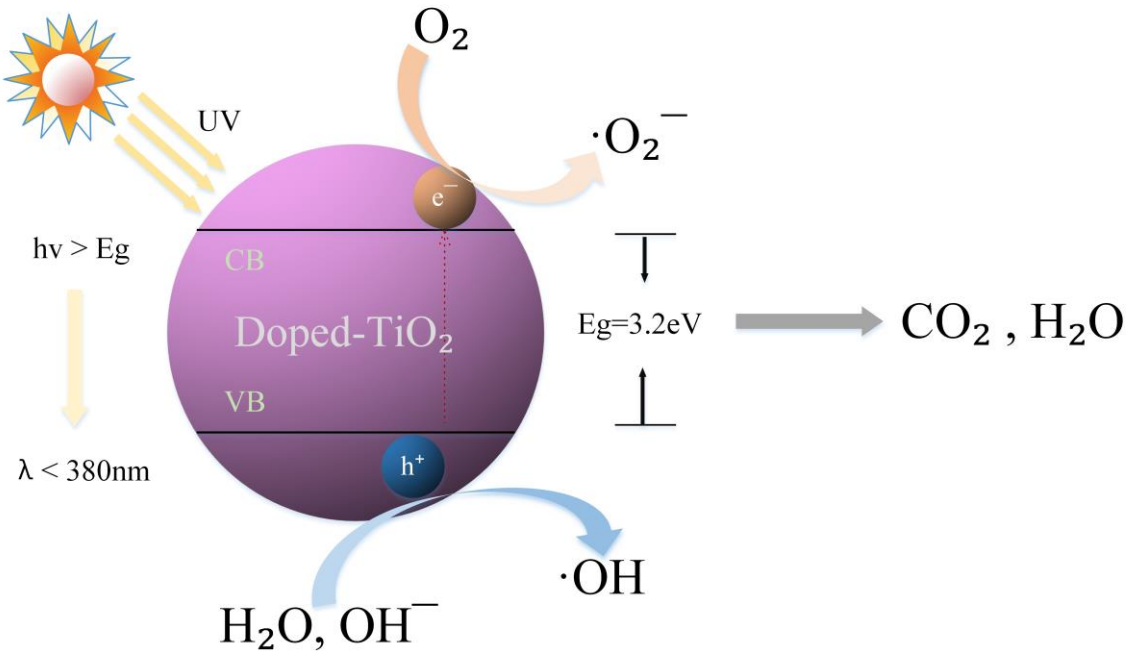

Figure 1. The schematic diagram of TiO_2 photocatalytic degradation of organic pollutants.

With the implementation of the material genome initiative (MGI), machine learning (ML), an artificial intelligence tool, has been successfully used in wider research communities, such as natural language processing [12], medical diagnosis [13], and biomedicine [14]. On the one hand, in the computational materials field including first-principle calculations, molecular dynamics and finite element simulation, researchers evaluated the photocatalytic degradation of ceftriaxone using heterogeneous O_3 / UV / Fe_3O_4@TiO_2 systems. They not only examine the effects of parameters (catalyst dosage, solution pH, initial concentration, and ozone concentration) on performance but also cover the photocatalytic oxidation processes using kinetic models [15]. S. Kang et al. developed a structured framework combining Neural Network potentials with evolutionary or random

searches to predict the structure of inorganic crystals and the program navigated configurational spaces10^2-10^3 times faster than the DFT-based method [16]. However, this method consumes considerable time and has some limitations. The application of ML could save calculation time to expand the space and time scale of the calculation system. On the other hand, the core statistical algorithm of machine learning has great capability for handling big data sets, thus it can correlate insightful information from the existing experimental database, explore the complex implicit relationship between various parameters, and establish an accurate prediction model through training and data mining. H. Masood et al. proposed an overall framework combining ML and domain knowledge to accelerate the development of solar photocatalysts [17]. Y. Zhang et al. developed a Gaussian regression model which can effectively estimate the E_g of TiO_2 and ZnO by statistically analyzing the relationship among the lattice parameters, grain size, surface area and energy bandgap [18-19]. The ultimate objective for ML model development is to predict experimental results without conducting experiments and these experimental conditions have never been encountered before.

In this paper, we studied the doped TiO_2 photocatalytic experimental data using four ML algorithms. By comparing the errors from the Linear Regression, Random Forest, XGBoost and LightGBM models, LightGBM is the most powerful model which was used to determine and rank the importance factor of the activity of the photocatalyst. Finally, we

also use the model to predict the degradation rate of photocatalysts under experimental conditions not previously encountered.

2. Materials and methods

2.1 Database construction

Initially, important factors affecting the photocatalytic activity of doped TiO_2 are obtained by referencing hundred of TiO_2 photocatalytic experiments from 2004-2021. Each data point contains nine experimental variables, i.e., dopant, dopant/Ti molar ratio, calcination temperature (°C), pollutant, catalyst/pollutant mass ratio, pH, experimental temperature (°C), light wavelength (nm) and illumination time (min). Dopants include non-metallic elements (such as C [20], F [21], I [22] and N [23]) and metal elements (such as Ag [24], Bi [25], Cd [26] and Ce [27]), respectively. Pollutants include methyl blue [28], phenol [29], methyl orange [30], benzoic acid [31], acid orange [32], etc. The dopant/Ti molar ratio ranges from zero to 93:5. The calcination temperature has the range from 400 to 900 °C. The mass ratio range is from 5:1 to 1000:1. The pH is in the range of 2 to 13. The experimental temperature is between 16 and 32 °C. The wavelength by illumination ranges from 254 to 600 nm. The illumination time is in the range of 5 to 480 min. All variables are summarized in Table 1. The output variable is the degradation rate, which is determined experimentally.

Table 1. The multiple input variables for training.

Variable	Range
Dopant	Ag, Bi, C, Ce, Cd, F, Fe, Ga, I, Mo, N, Ni, S
Dopant/Ti mole ratio	0 - 93:5
Calcination temperature	400 - 900
Pollution	Methylene Blue (MB), Phenol, Rhodamine B (RhB), Methyl Orange (MO), Methyl Red (MR), Acid Orange (AO)
Catalyst/Pollutant mass ratio	5:1 - 1000:1
pH	2 - 13
Experimental temperature	16 - 32
Light wavelength	254 - 600
Illumination time	5 - 480

For typical doped TiO_2 photocatalytic experiments, different types of modified TiO_2 were prepared by the sol-gel method under adjusted calcination temperature and the portion of dopants and TiO_2. Next, a certain amount of doped TiO_2 photocatalyst was added into the pollutant solution with the different pH values. After stirring in the dark environment for a while to reach the adsorption equilibrium, then the suspension was illuminated by light sources with different wavelengths. Within a given time interval, a small portion of the suspension was removed, and all solid particles were filtered out. Finally, the concentration of the contaminated solution was measured by a UV-vis spectrophotometer. The degradation rate can be calculated by following equation:

$$\eta = \frac{C_0\text{-}C}{C} \times 100\% \quad (6)$$

The η represents the degradation rate, C_0 is the initial concentration of the contaminated solution before illumination, and C is the solution concentration after treatment at any time t.

Finally, in total of 760 data points from 28 publications (Supplementary Materials) are used for model training.

2.2 Model introduction

The basic procedure of ML and the associated models are explained as follows: the programming platform is Jupyter NoteBook, a website-based interactive computing environment. Jupyter is an integrated development environment (IDE) for Python in Anaconda. Three python libraries were used, including NumPy, Pandas and Scikit-learn. As shown in Figure 2, the data requires pre-processing before further modeling. As ML models are based on mathematical functions; two variables (dopant and pollutant type) cannot be imported and trained directly, a LabelEncoder was thus used to encode these two variables. The rest variables are readable as numerical data. The seaborn pairplot function is employed to analyze the correlation between nine input variables in Section 3.1. Then, the datasets was split into the training set and the testing set with a ratio of 7:3. Moreover, the model reliability could be improved via the K-fold cross-validation (CV) method to avoid overfitting and underfitting. K-fold CV could randomly divide the original training set into k parts, selecting (k-1) as the training set whilst the remaining 1 as the validation set. The CV was repeated K times, and thus the average accuracy of K times is taken as the evaluation index for the final model. The selection of the K value can be flexible according to the actual situation, and 10 CVs were adopted in this study.

In the experiments, we use four models: Linear Regression (LR), Random Forest (RF), XGBoost (XGB) and LightGBM (LGB). Through the ten-fold CV method, we find the model with the best generalization performance and retrain it on original training set, and the testing set is used to make the final evaluation of the model performance. The details of the four models are described in the Supplementary Materials.

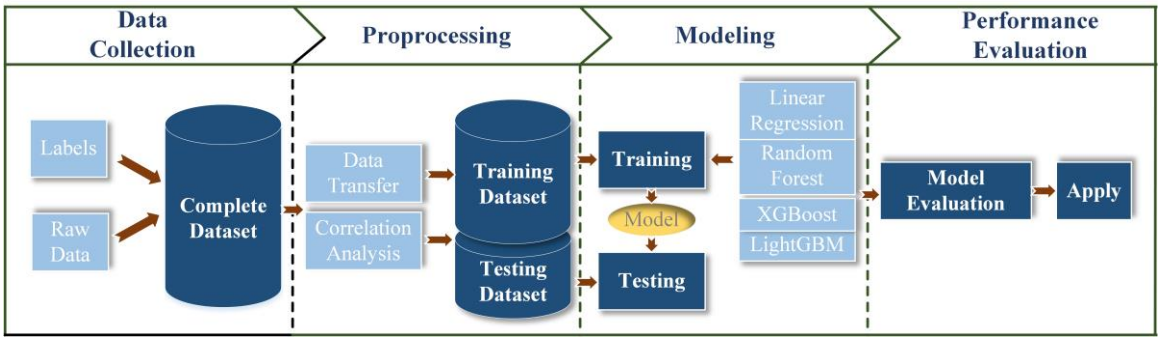

Figure 2. The flow chart of machine learning.

3. Results and discussion

3.1 Correlation analysis

Correlation analysis is analyzing two or more variables to measure the closeness between these variables. If there is a large value of linear correlation between the two input variables, it may lead to the prediction failure. Therefore, one should analyze the correlation of input variables before establishing the prediction model. If there are two input variables with a large linear correlation, either one of them will be removed or the sample size needs increase.

We evaluated the correlations among all variables (i.e., dopant type, dopant/Ti molar ratio, calcination temperature, pollutant type, catalyst/pollutant mass ratio, pH,

experimental temperature, light wavelength, catalytic time). It concludes that there was barely linear relationship between any pair of variables. The relationships among the variables are illustrated in Figure 3 and Figure 4.

The pairplot in Figure 3 shows the correlation between the four variables (dopant, dopant/Ti molar ratio, calcination temperatures, and catalyst/pollutant mass ratio) regarding the degradation rates. Here, the degradation rate is divided into five levels marked in blue (0-20), orange (20-40), green (40-60), red (60-80) and purple (80-100). The diagonal subgraphs represent the data distribution of each variable in the database, the x-axis represents the range of variables, and the y-axis represents the distribution density of variables. For example, the first subfigure in the1st row of Figure 3 shows the distribution of calcination temperature at different levels against the degradation rates. The 2nd to 4th subfigures in the first row shows the function of dopant, dopant/Ti molar ratio, and catalyst/pollutant mass ratio (x-axis) against the temperature (y-axis). Similarly, the remaining subgraphs are the correlation diagrams in the other three rows.

There is no linear relationship between the four input variables. For each non-diagonal subfigure, the straight line is the linear model of y ~ x fitted with data and the shaded area represents the 95% confidence interval. The probability that the invisible data will fall around the fitting line. While the linear model does not apply to the nonlinear relationship. Instead, a nonlinear model should be selected. The correlations between all variables are presented in Supplemental Materials (Figure S1).

In Figure 4, a 9 x 9 heatmap is plotted to show the correlation matrix. Each number in the sub-square represents the Pearson correlation coefficient of two variables. The darker the color, the value of the correlation coefficient is closer to 1. The higher value in each sub-square also indicates the two variables have a stronger correlation level. There is no linear relationship between each variable as shown in Figure 4, confirming the previous conclusions (Figure 3). This attributes to the complex catalytic mechanism, it is thus difficult to find a certain variable that greatly affects the catalytic activity according to Figures 3 and 4. But it helps the subsequent model selection, the effect of the nine variables on the degradation rate is described in section 3.3 and analyzed by a nonlinear model.

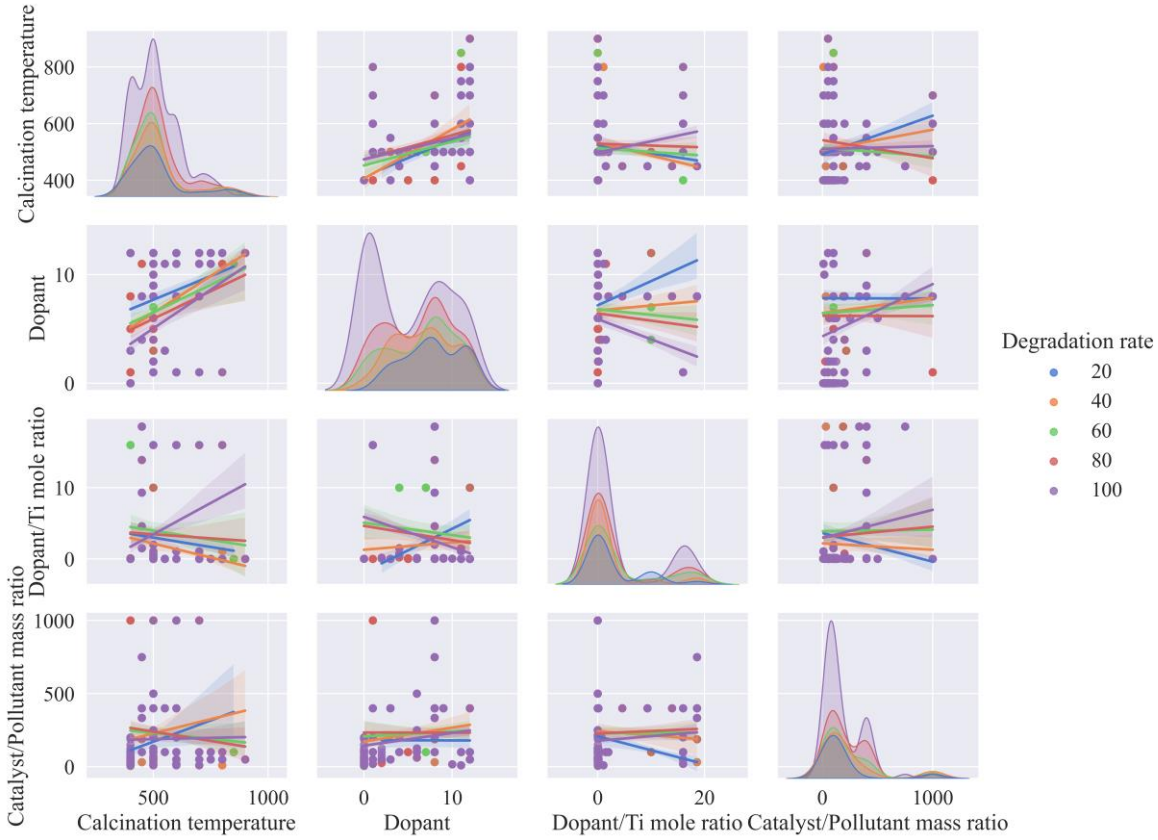

Figure 3. The correlation between calcination temperature, dopant, dopant/Ti mole ratio and

catalyst/pollutant mass ratio.

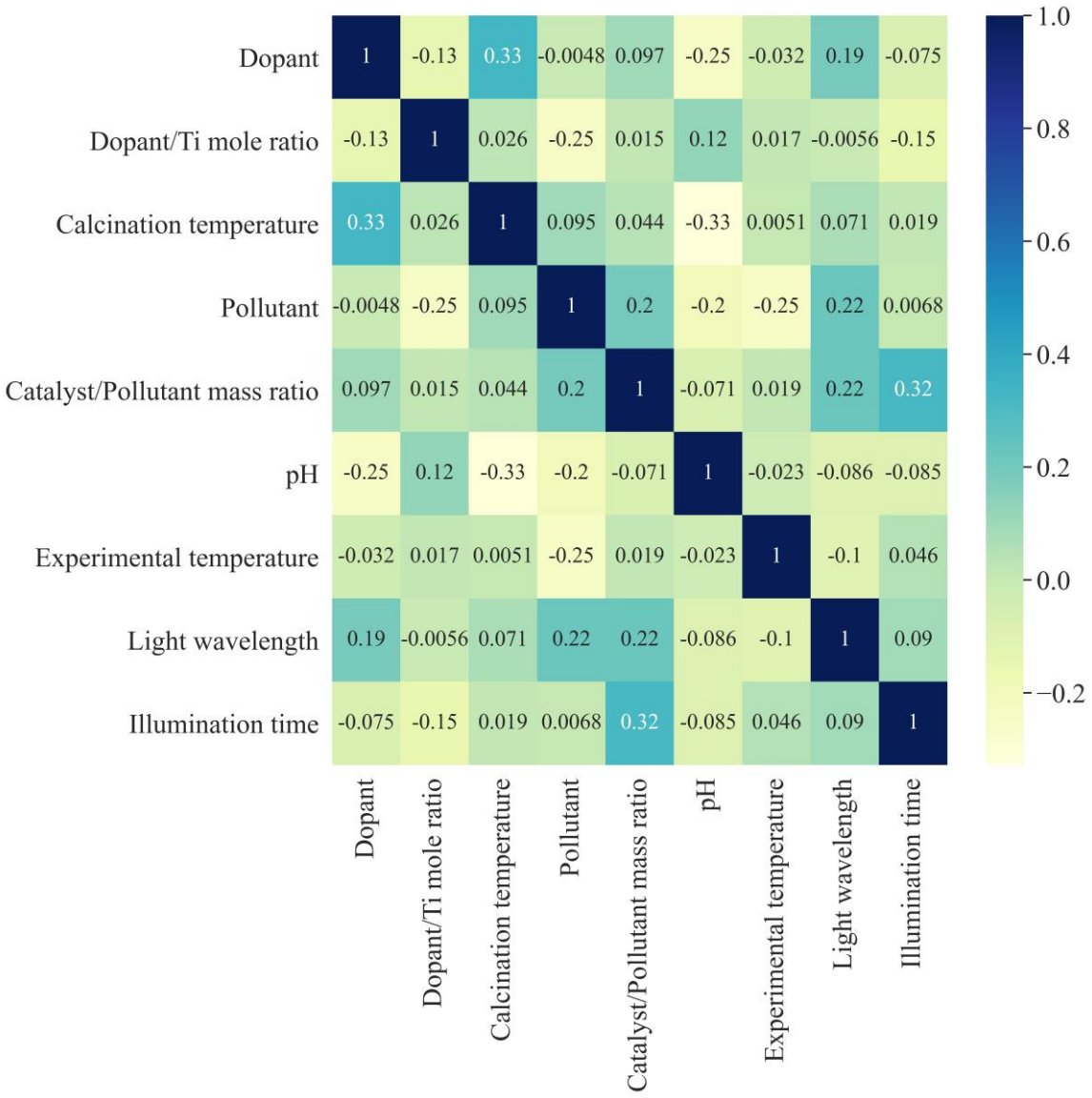

Figure 4. The heat map showing the correlation between nine input variables

3.2 Model accuracy

In this study, we use four indicators to evaluate the proposed four models (i.e., LR, RF, XGB, and LGB). Three errors, namely the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), are used to calculate prediction errors, and the coefficient of determination (R^2) is used to measure the correlation level. MSE value is the square of the difference between the predicted value and the real value. The smaller the value of MSE, the better the accuracy of the prediction model. RMSE (as a standard deviation of the residual) reveals the error between the predicted value and the real value. When R^2 (maximum value is 1) is applied to the test set, the value is equal to the variance of external interpretation, which can be used to determine the model quality. When R^2 is closer to 1, the better the fitting degree of the model. Here, both RMSE and R^2 are selected as the prime indicators to choose the best prediction model. The low RMSE and high R^2 score will indicate a good fitting model.

In Table 2, the LR showed the worst performance, with an RMSE of 0.762 and an R^2 score of 0.048. This also verifies the prior results in Section 3.1, suggesting the linear model cannot properly fit the relationship between various variables. Whilst those typical nonlinear regression models (i.e., RF, XGB, LGB) have significant improvement regarding both RMSE and R^2 values. This is supported by having a relatively low RMSE (below 0.45) and high R^2 (beyond 0.8). In particular, the LGB is the most stable and reliable model, with an RMSE of 0.194 and an R^2 score of 0.899. Therefore, modeling photocatalytic degradation by the LGB model yields the best generalization capability, prediction

performance and the strongest robustness. Figure 5 visualizes the average error of the four models, where each subgraph shows the relationship between the experimental value (Y-axis) and the predicted value (x-axis). The dotted line in this figure represents the perfect prediction line when y = x. The synthesis of Table 2 and Figure 5 shows that the predicted value of the LGB model can better simulate the real value. Hence, the LGB model is accepted as the final prediction model.

Table 2. The prediction performance comparison of LR, RF, XGB and LGB models according to MAE, MSE, RMSE and R^2.

	Linear Regression	**Random Forest**	**XGBoost**	**LightGBM**
MAE	0.513 ± 0.104	0.235 ± 0.062	0.145 ± 0.103	0.116 ± 0.065
MSE	0.601 ± 0.237	0.180 ± 0.126	0.086 ± 0.034	0.043 ± 0.031
RMSE	0.762 ± 0.401	0.417 ± 0.148	0.293 ± 0.136	0.194 ± 0.101
R^2	0.048 ± 0.014	0.805 ± 0.035	0.884 ± 0.024	0.898 ± 0.012

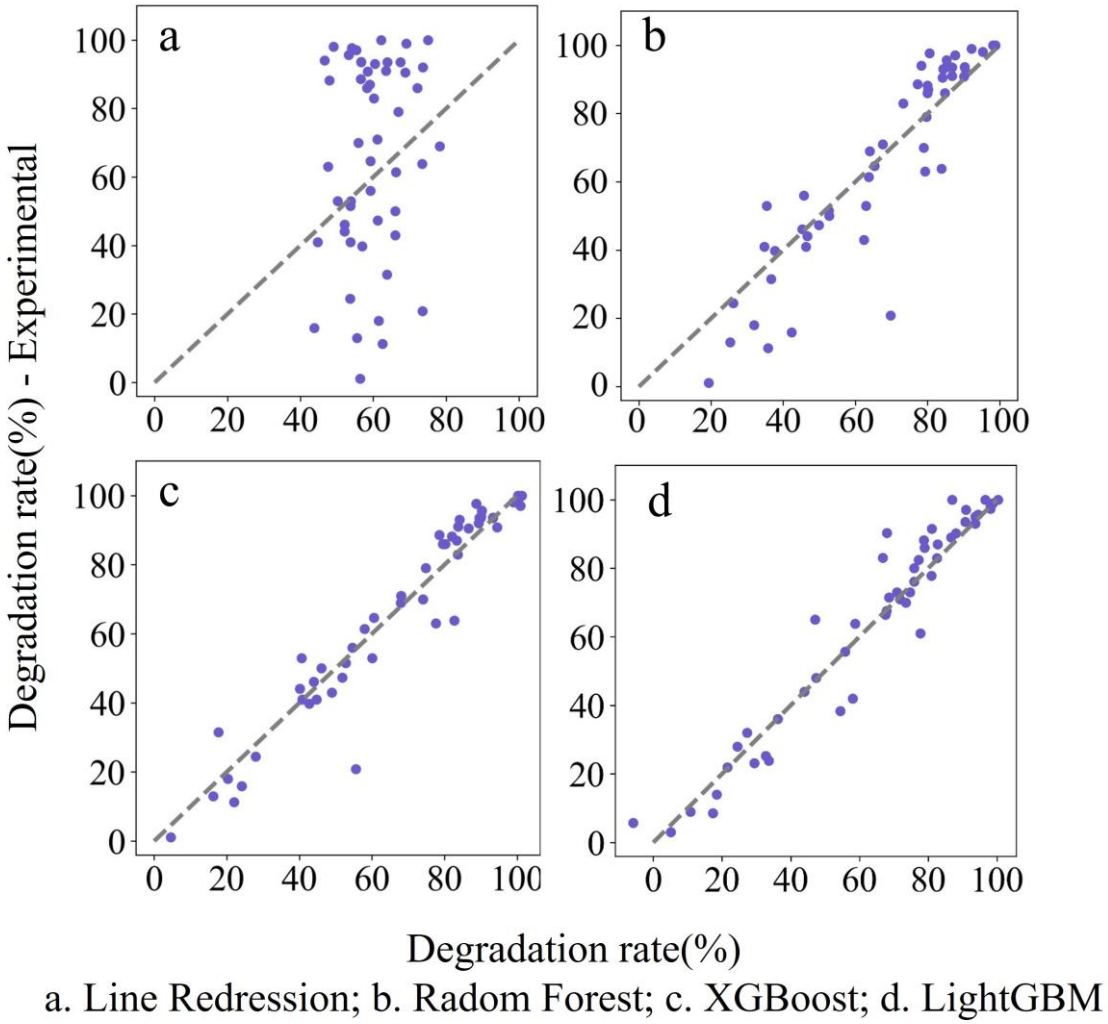

Figure 5. The comparison of average errors among four models

LGB model is used to re-train the original training set resulting in the R^2 score of 0.928. All parameters used for ML models are presented in the Supplementary Materials. The predicted value equals to the actual value when discrete data point is featuring on the y=x line. The proximity between the discrete data points and the y = x line shows the reliable prediction via the LGB model. Moreover, the difference between the predicted value and the observed value under the regression model must be random and unpredictable;

that is, there should be no interpretable and predictable information in the error. The residual diagram has dual functions as it can be used to intuitively observe the difference between the prediction and actual value as well as model quality evaluation. Figure 7 shows the relationship between the predicted value (x-axis) and residual value (y-axis). Theoretically, the best model (errorless) has zero residual, which is almost impossible in practical applications. However, for a good model, the errors could be randomly distributed. It is expected to see the residual values fluctuate near the horizontal line (y = 0). Figure 7 shows that the testing data points are distributed next to the y=0 line with most residuals < 20 %, affirming the validity of the LGB model.

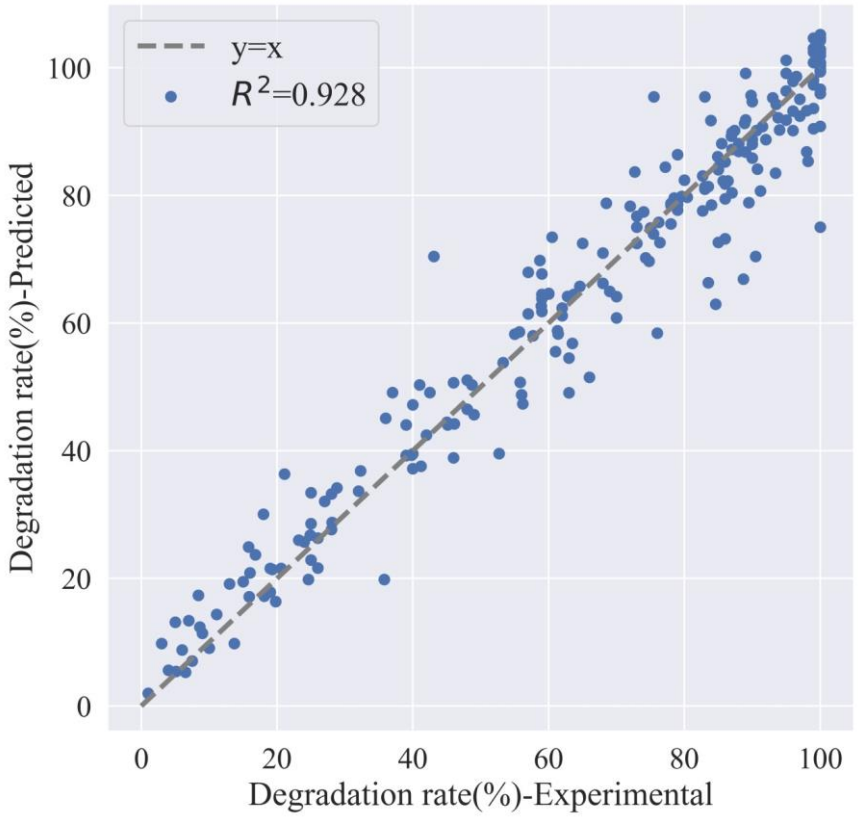

Figure 6. The error of LGB prediction model after retraining in the original training set.

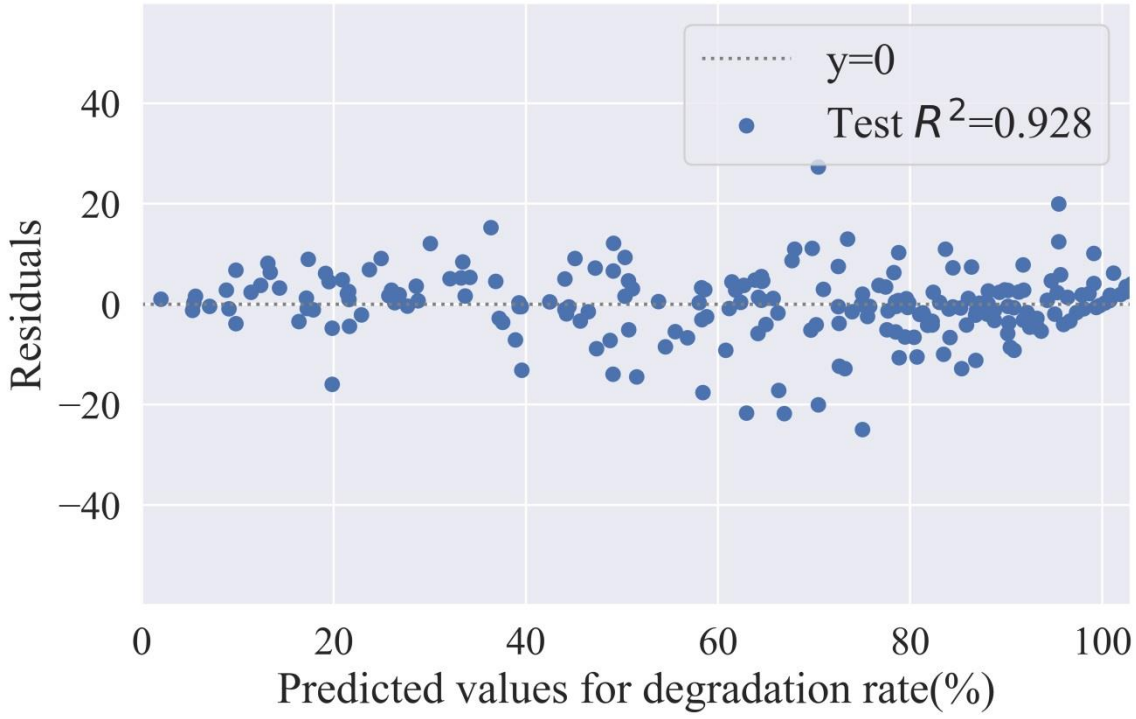

Figure 7. The residual error of LGB prediction model after retraining in the original training set.

3.3 The effect of the influencing factors on the degradation rate

The importance of each input variable is investigated. The importance scoring of a feature is a means of scoring input features, revealing the relative importance of each feature for LGB model when making predictions upon how useful each input feature is in predicting the target variables.

The significance of a feature is to reduce the uncertainty of the prediction target, and the feature that can reduce more uncertainty is crucial. This helps to achieve a better understanding of the effects of the experimental variables on the performance of the photocatalytic degradation process, thus offering guidance for practical design in real applications. Herein, feature importance values are calculated based on the number of times

that a feature is used in all trees. As shown in Figure 8, the importance of the nine features is ranked for the following analysis.

In Figure 8, the illumination time (score of ca. 2750) is the most influential variable for photocatalytic degradation experiments. Generally, the photocatalytic degradation rate gradually increases with prolonged illumination time [33]. The reaction rate of photocatalytic oxidation can be described by the Langmuir-Hinshelwood kinetic equation [34]. With the progress of the reaction, both the concentration of organic pollutants and reaction rate will decline, thus the degradation rate will not further increase and reach a plateau after a certain period. C. Sahoo et al. [35] studied the use of Ag ion-doped TiO_2 as a photocatalyst to degrade textile wastewater. The results show that the degradation rate increase in the first 45 minutes before degradation, and then the reaction rate slows down from 45 to 150 minutes. After 150 minutes, prolonging the illumination time does not increase the degradation rate at all. C. H. Chiou et al. [36] used Pr-doped TiO_2 nanoparticles to photocatalyze the degradation of phenol, and the degradation rate peaked at 90 minutes. These two cases suggest that doping elements, organic pollutants and other experimental variables have different illumination times to achieve their best degradation efficiency. Therefore, before performing the photocatalytic degradation experiment, the issue we must tackle is to determine the illumination time for the specific experiment condition.

TiO_2 semiconductors have stable properties as the electrons are not easily excited. Due to this characteristic, various dopants have been introduced into the TiO_2 photocatalyst to

enhance its photocatalytic efficiency. A dopant will be beneficial to capture electrons which reacts with water molecules, forming oxidative radicals to degrade organic pollutants. However, when the additive dopant accumulates to an excessive concentration, the dopant will negatively prevent TiO_2 from capturing photons by (a) reducing the active surface area of the photocatalyst and (b) enhancing the photoinduced charge carrier recombination by narrowing the space charge region [37-38]. J. Shi et al. [39] studied the concentration of samarium as a dopant in titania catalysts for photocatalytic degradation of methyl orange. They observed that upon adding dopant concentrations of 0.05-0.1 mol% to titania, the photodegradation efficiency increase with increasing concentration and reach to the maximum. The degradation efficiency compared with that of pure TiO_2 is improved by restraining the crystal size (with the larger surface areas), helping the transportation and exchange of organic matter, facilitating the movement of photogenerated support to the surface of the photocatalyst and preventing the recombination of support. However, it is noticeable that a significant reduction in the degradation efficiency of the photocatalyst. When the doped ion concentration is too high, the space charge region becomes very narrow, and the penetration depth of light into TiO_2 greatly exceeds the space charge layer, thus the recombination of the electron-hole pairs in the semiconductor becomes easier. Based on the above descriptions and current model, it is reasonable to classify the dopant/Ti mole ratio catalyst as the secondary important factor for TiO_2 degradation.

The catalyst/pollutant mass ratio includes two experimental variables: the amount of catalyst and the initial concentration of pollutant solution. Generally, when the catalyst dosage increases, the degradation rate of photocatalytic degradation pollutants increases. This attributes to the rise of the catalyst surface area, which increases the contact probability between organic matter and the catalyst, thus promoting the progress of the catalytic reaction. However, when the amount of catalyst reaches a certain limit, the photodegradation rate of pollutants will decline. On the one hand, this may be attributed to a higher dosage of catalyst. The catalysts overlap each other resulting in a higher thickness that might block the transmission depth of illumination, thus leading to the reduction of the photocatalytic effect due to light scattering. Another reason might be the agglomeration of nanoparticles at high concentrations in the solution, limiting the number of active surface sites available for exposure [40]. Similarly, if the initial concentration of pollutant solution exceeds the limit, the photodegradation rate will also decrease. Because the pollutants completely cover the surface of the photocatalyst, the number of photons reaching the surface of the catalyst and the number of hydroxyl ion free radicals and positive holes are reduced [41]. C. Sahooet et al. used silver doping TiO_2 as a photocatalyst to study the effect of the initial dye concentration, which is in the range of 10-50 ppm, on the photodegradation of methyl red [42]. Hence, the catalyst/pollutant mass ratio is regarded as the 3rd important factor.

In most cases, anatase TiO_2 has higher photocatalytic activity than rutile TiO_2. The reason is that the anatase phase lattice contains more defects and dislocations, which can produce more oxygen vacancies to capture electrons, resulting in easier separation of photogenerated electrons and holes. The calcination temperature has a great influence on the crystal shape and surface morphology of the catalyst. As a result, it is listed as the fourth most important variable affecting the degradation rate. M. Hamadanian et al. [43] when studying the effect of calcination temperature on the photocatalytic activity of S-doped TiO_2, found that when the calcination temperature was 500 °C, S-doped TiO_2 only included the anatase phase and at this time, the photodegradation rate was the highest. Then, with increasing temperature, the degradation rate began to decline, and the formation of the rutile phase started at 750 °C. This is because with increasing calcination temperature, sintering occurs, and the surface area decreases, resulting in a reduction in catalytic activity.

The light wavelength, dopant and pH have a similar influence on the degradation, with their absolute scores close to ca. 1000. As the core part of photocatalytic technology, the light source is an independent factor in the photocatalytic reaction system and a necessary condition to stimulate the photocatalytic degradation. Previous experiments show that the light wavelength affects the photocatalytic reaction under different light sources, and the wavelength applied to the photoactivation has an impact on the dye degradation speed; shorter wavelengths lead to faster degradation, and vice versa [44]. The dopant additives widen the visible light range and thus improve the photocatalytic efficiency, but not all

doped elements can do so. Only suitable semiconductor materials could influence the photoelectrochemical properties of TiO_2 and change the band gap of the photocatalyst [45], and the mechanism of catalytic activity is also different. Studies show that after 10 hours of doping of metal and nonmetallic elements, such as C, Fe, N, and Zn into TiO_2, C-TiO_2 has a high photocatalytic activity under visible light, mainly attributed to small particle size, high specific surface area, and good light absorption and photocatalytic activity within the visible light region [46]. Therefore, the selection of dopants determines photocatalytic degradation. The change in pH value will alter the adsorption of OH^-, H^+ on the photocatalyst surface and the surface charge of TiO_2, influence the adsorption of reactants on the TiO_2 surface and finally affect the photocatalytic reaction rate. In addition, different structures of reactants have different adsorption capacities for ·OH on the photocatalyst surface, resulting in different initial pH values for the degradation of different reactants [47-48]. For example, if pH goes up, the efficiency of photocatalytic degradation of rhodamine B gradually increases as the pH manipulates the charge of rhodamine B [49]. When the pH < 4, rhodamine B exists in the form of cations, the adsorption on the catalyst surface becomes difficult since the action of electrostatic repulsion, as a result, the degradation efficiency will be reduced. When pH > 4, rhodamine B exists in the form of zwitterions. Due to the action of electrostatic attraction, some molecules are adsorbed on the catalyst surface, and thus degradation efficiency increases.

The type of pollutants and experimental temperature, as the two least important features, also have an impact on photocatalytic activity. The worst example is that the photocatalytic degradation rate of monoazo dyes is higher than that of dyes with antraquinone structures; the presence of methyl and chloro groups in the dye molecule slightly lowers the efficiency, whilst a nitrite group acts oppositely [50]. Since most photocatalytic experiments are performed at near room temperature, the experimental temperature has little effect on the photocatalytic reaction. However, some researchers found that the photocatalytic degradation rate could go up with the increase in temperature [51].

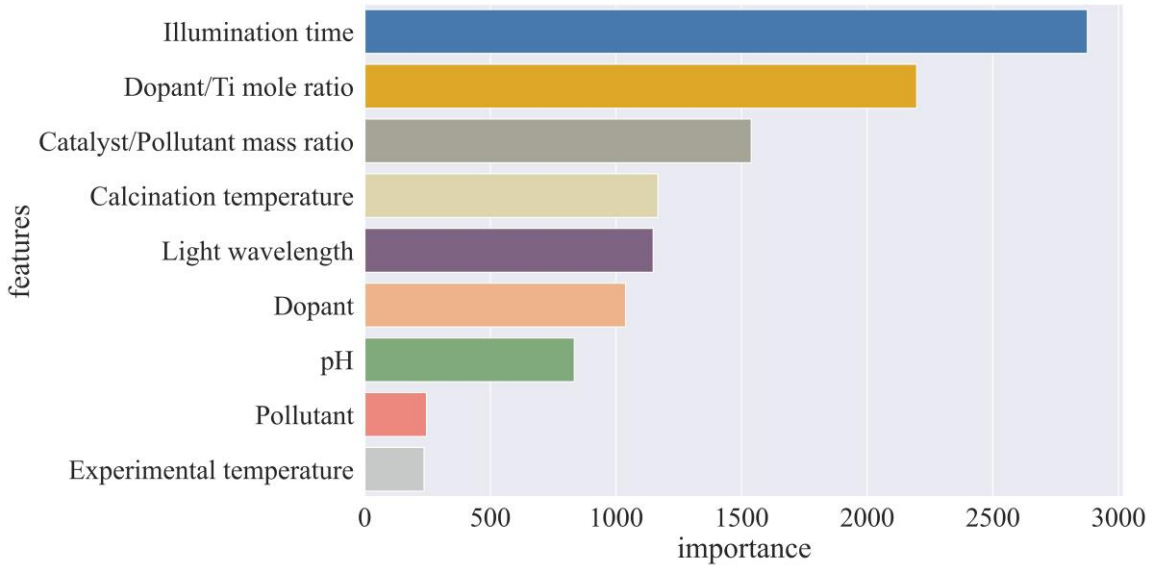

Figure 8. The chart of importance analysis based on nine influencing factors.

3.4 Model Verification

In this section, we look at four successful cases to demonstrate the good prediction in new materials by our pre-trained model. Fe-ions can be doped into TiO_2 to reduce the

recombination of photogenerated electrons and holes and thus enhance the photocatalytic activity of TiO_2 due to its half-filled d-electronic configuration and identical ionic radius to Ti^{4+} [52-53]. Both pure TiO_2 and Fe-doped TiO_2 photocatalysts are prepared at a calcination temperature of 400 °C. They are pure crystals and have an anatase phase. Fe-doped TiO_2 can effectively degrade methylene blue under visible light irradiation. In Figure 9, the Fe-doped TiO_2 sample (70 %) demonstrates a greater degradation rate than that in pure TiO_2 (35%), nearly doubled. The trained LGB model is employed to predict the degradation rate of Fe-doped TiO_2 under the same experimental conditions as the input variables. Notably, these input variables are excluded from our database and have not been trained. The green line (Figure 9) depicts the predicted profile, suggesting the max degradation rate (ca. 65 %) is slightly lower than that achieved by the experiment (70 %). Although these two curves are not overlapped, the trend of both curves develops quite identical, and the deviation is within an acceptable error range. The result shows the photocatalytic activity of TiO_2 will be significantly improved by the Fe ions additives.

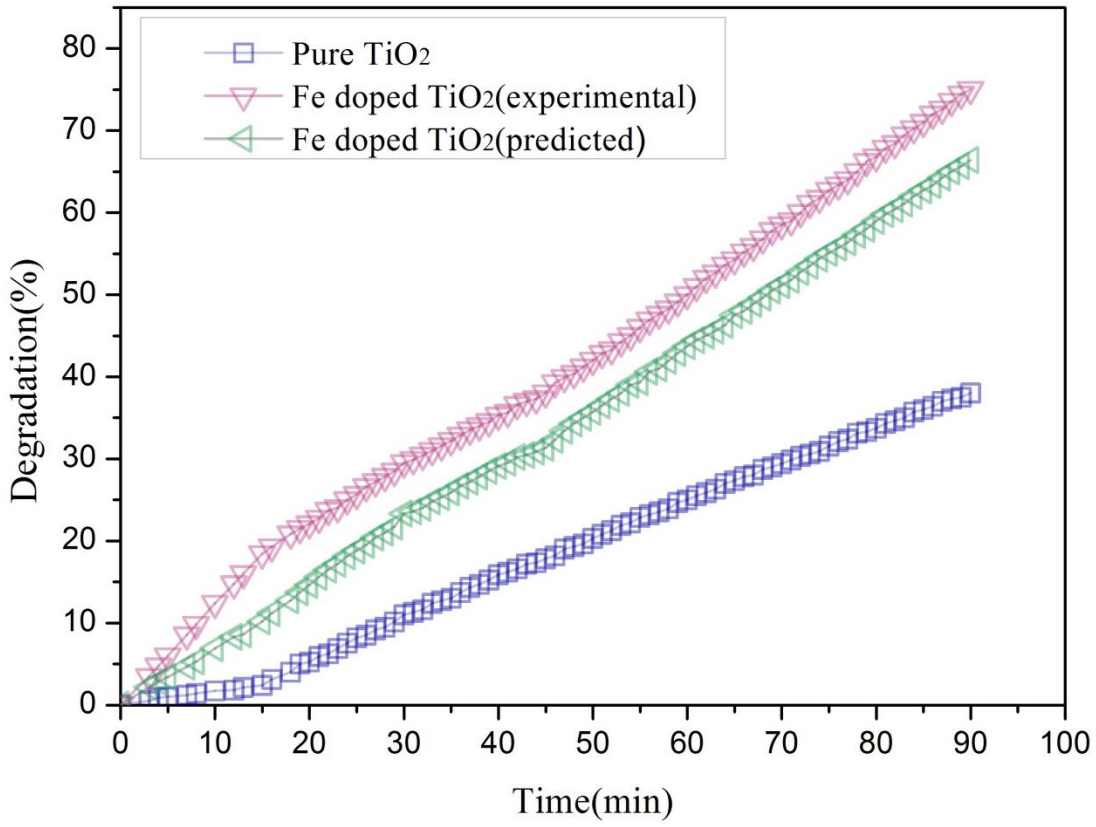

Figure 9. Concentration of the degradation rate as a function of irradiation time in the presence of the pure and Fe-doped TiO_2 photocatalyst.

Similarly, another three datasets are tested by our LGB model [54-57]. Figure 10(a) shows the Cd-doped TiO_2 case study under simulated sunlight indoors. With increasing dopant loading, the predicted curve by the pre-trained LGB model (blue line) develops a similar trend to the experimental curve (red line) under identical conditions. Figure 10(b) and Figure 10(c) illustrate the Ag-doped TiO_2 with the change of catalyst/pollutant mass ratio and C-doped TiO_2 with the change of calcination temperature, respectively. In these two cases, the predicted values (blue line) are almost superimposed on the experimental data (green line), suggesting a high accuracy of prediction.

Based on presented case studies above, our trained LGB model could provide the accurate prediction on the degradation rate of doped TiO_2. It is envisaged that the prediction accuracy will be further improved with the increase of the dopant database. In fact, it is extremely useful to obtain some reliable ML predictions to guide the actual experiments.

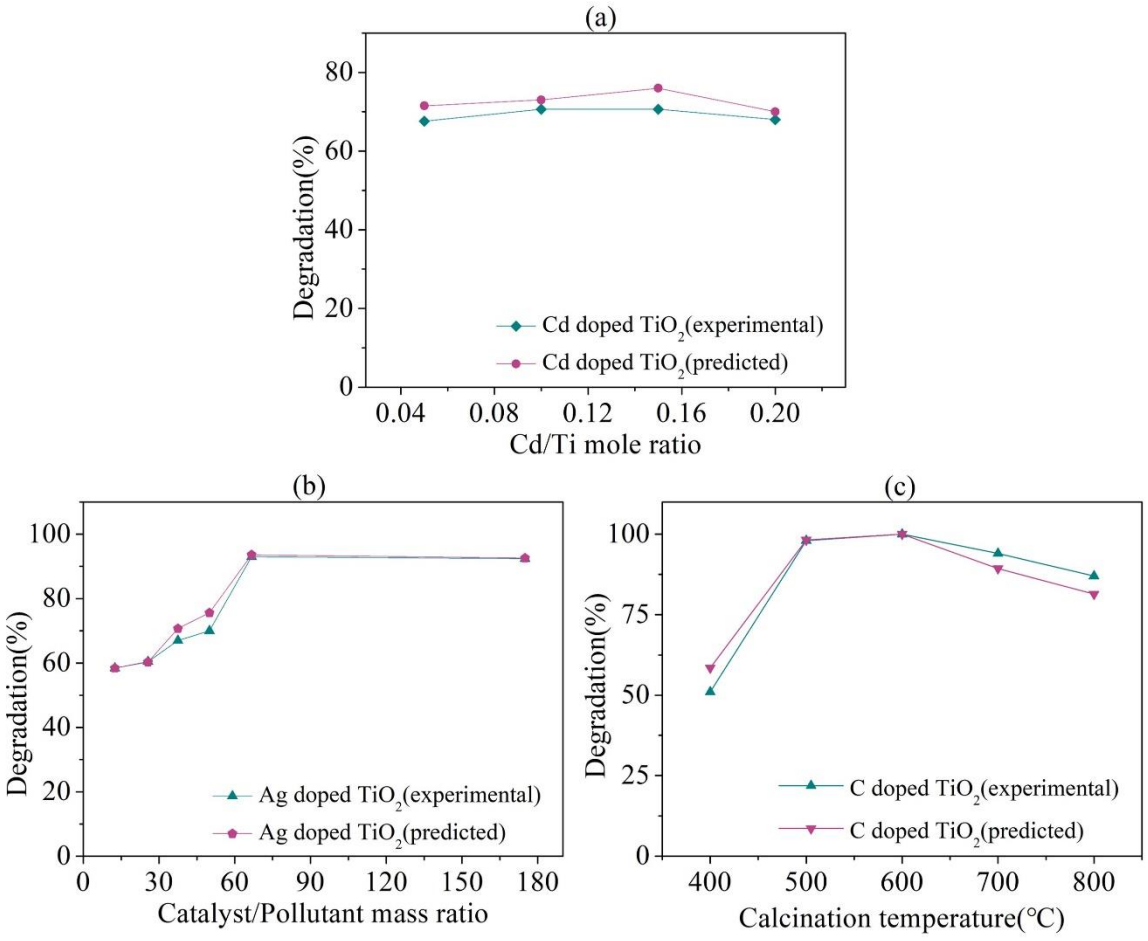

Figure 10. Influence of Cd/Ti mole ratio, catalyst/pollutant mass ratio, calcination temperature on photocatalytic performance of doped TiO_2 nanophotocatalyst: (a) calcination temperature: 500 °C, pollutant type: acid orange, catalyst/pollutant mass ratio: 50; (b) Ag/Ti mole ratio: 0.0101, calcination temperature: 500 °C, pollutant type: methylene blue; (c) C/Ti mole ratio: 16, pollutant type: methylene blue, catalyst/pollutant mass ratio: 100.

4. Conclusions

In this study, the photocatalytic degradation rate of organic wastewater pollutants in the presence of various doped TiO_2 photocatalysts has been comprehensively simulated using a machine learning approach. The selected data includes published 760 data points. By comparing the errors of the linear regression, random forest, XGBoost, and LightGBM (LGB) models, the LGB model with the best credibility is selected to make the further prediction. The application of ML provides a feasible route to rank the importance of experimental variables that affects catalytic activity in terms of the degradation of doped TiO_2. The LightGBM model suggests the following influential sequence: illumination time > dopant/Ti molar ratio > catalyst/pollutant mass ratio > calcination temperature > light wavelength > dopant > pH > pollutant > experimental temperature. In addition, another four independent cases, including the Fe, Cd, C and Ag doped TiO_2, are used to validate the model accuracy by comparing the experimental data with predicted values. Therefore, we envisage that the ML model-driven prediction can be used as guidance before actual doped-TiO_2 experiments. Moreover, the volume and dimension of the current data pool is expected to expand via future high-throughput experiments, this will allow ML prediction to achieve higher accuracy than the current status and ultimately reduce the cost. This pre-trained ML model could also be applied to other photocatalysts, greatly accelerating the development of photocatalytic degradation technology.

Supplemental Materials

This section includes the details of the four models (Linear Regression, Random Forest, XGBoost and LightGBM), all datasets, the relationship between nine variables, and the parameters of the predicted model.

Acknowledgements

We acknowledged the financial supports from the Foundation for High-level Talents in Higher Education of Hubei (163083), and the Integration of Industry, Education and Research of Science and Technology Department of Hubei Province (CXYH2019000301). WD thanks the Enrichment Award from the Alan Turing Institute (UK).

Supporting Information

All our scripts and data are available on github ()

References

[1] J. Tollefson, How green is my future? Nature. 473 (2011) 134-135. https://doi.org/10.1038/473134a

[2] L. Carlsen, R. Bruggemann, Y. Sailaukhanuly, Application of selected partial order tools to analyze fate and toxicity indicators of environmentally hazardous chemicals, Ecol. Indic. 29 (2013) 191-202. https://doi.org/10.1016/j.ecolind.2012.12.028

[3] V. Katheresan, J. Kansedo, S. Y. Lau, Efficiency of various recent wastewater dye removal methods: A review, J. Environ. Chem. Eng. 6 (4) (2018) 4676-4697. https://doi.org/10.1016/j.jece.2018.06.060

[4] S. Y. LeeQ, S. J. Park, TiO_2 photocatalyst for water treatment applications, J. Ind. Eng. Chem. 19 (6) (2013) 1761-1769. https://doi.org/10.1016/j.jiec.2013.07.012

[5] F. Spadavecchia, M. Ceotto, L. L. Presti, C. Aieta, I. Biraghi, D. Meroni, S. Ardizzone, G. Cappelletti, Second generation nitrogen doped titania nanoparticles: A comprehen-sive electronic and microstructural picture, Chin. J. Chem. 32 (12) (2014) 1195-1213. https://doi.org/10.1002/cjoc.201400502

[6] A. Fujishima, T. N. Rao, D. A. Tryk, Titanium dioxide photocatalysis, J. Photochem. Photobiol. C. 1 (1) (2000) 1-21. https://doi.org/10.1016/S1389-5567(00)00002-2

[7] W. Zhao, W. Ma, C. Chen, J. Zhao, Z. Shuai, Efficient Degradation of Toxic Organic Pollutants with $Ni_2O_3/TiO_{2-x}B_x$ under Visible Irradiation, J. Am. Chem. Soc. 126 (15) (2004) 4782-4783. https://doi.org/10.1021/ja0396753

[8] J. T. Park, D. J. Kim, D. H. Kim, J. H. Kim, A facile graft polymerization approach to N-doped TiO2 heterostructures with enhanced visible-light photocatalytic activity, Mater. Lett. 202 (2017) 66-69. https://doi.org/10.1016/j.matlet.2017.05.070

[9] M. Khairy, W. Zakaria, Effect of metal-doping of TiO_2 nanoparticles on their photocatalytic activities toward removal of organic dyes, Egypt. J. Petrol. 23 (4) (2014) 419-426. https://doi.org/10.1016/j.ejpe.2014.09.010

[10] D. Chen, Y. Cheng, N. Zhou, P. Chen, Y. Wang, K. Li, S. Huo, P. Chen, P. Peng, R. Zhang, L. Wang, H. Liu, Y. Liu, R.Ruan, Photocatalytic degradation of organic pollutants using TiO_2-based photocatalysts: Areview, J. Clean. Prod. 268 (2020) 121725. https://doi.org/10.1016/j.jclepro.2020.121725

[11] M. R. Eskandarian, H. Choi, M. Fazlia, M. H. Rasoulifard, Effect of UV-LED wavelengths on direct photolytic and TiO_2 photocatalytic degradation of emerging contaminants in water, Chem. Eng. J. 300 (2016) 414-422. https://doi.org/10.1016/j.cej.2016.05.049

[12] H. Liu, C. Xu, J. Liang, Dependency distance: A new perspective on syntactic patterns in natural languages, Phys. Life. Rev. 21 (2017) 171-193. https://doi.org/10.1016/j.plrev.2017.03.002

[13] A. Criminisi, Machine learning for medical images analysis, Med. Image. Anal. 33 (2016) 91-93. https://doi.org/10.1016/j.media.2016.06.002

[14] S. Ekins, A. C. Puhl, K. M. Zorn, T. R. Lane, D. P. Russo, J. J. Klein, A. J. Hickey, A. M. Clark, Exploiting machine learning for end-to-end drug discovery and development, Nat. Mater. 18 (5) (2019) 435-441. https://doi.org/10.1038/s41563-019-0338-z

[15] S. Y. Hashemia, M. Y. Badi, H. Pasalari, A. Azarid, H. Arfaeinia, A. Kiani, Degradation of Ceftriaxone from aquatic solution using a heterogeneous and reusable O_3/UV/ Fe_3O_4@TiO_2 systems: operational factors, kinetics and mineralisation, Int. J. Environ. An. Ch. (2020) 1-7. https://doi.org/10.1080/03067319.2020.1817909

[16] S. Kang, W. Jeong, C. Hong, S. Hwang, Y. Yoon, S. Han, Accelerated identification of equilibrium structures of multicomponent inorganic crystals using machine learning potentials, NPJ. Comput. Mater. 8 (2022) 108. https://doi.org/10.1038/s41524-022-00792-w

[17] H. Masood, C. Y. Toe, W. Y. Teoh, V. Sethu, R. Amal, Machine Learning for Accelerated Discovery of Solar Photocatalysts, ACS. Catal. 9 (2019) 11774−11787. https://doi.org/10.1021/acscatal.9b02531

[18] Y. Zhang, X. Xu, Machine Learning Band Gaps of Doped-TiO_2 Photocatalysts from Structural and Morphological Parameters, ACS Omega. 5 (25) (2020) 15344-15352. https://doi.org/10.1021/acsomega.0c01438

[19] Y. Zhang, X. Xu, Machine learning optical band gaps of doped-ZnO films. Optik, 217 (2020) 164808. https://doi.org/10.1016/j.ijleo.2020.164808

[20] F. Dong, S. Guo, H. Wang, X. Li, Z. Wu, Enhancement of the Visible Light Photocatalytic Activity of C-Doped TiO_2 Nanomaterials Prepared by a Green Synthetic Approach, J. Phys. Chem. 115 (27) (2011) 13285-13292. https://doi.org/10.1021/jp111916q

[21] W. Yu, X. Liu, L. Pan, J. Li, J. Liu, J. Zhang, P. Li, C. Chen, Z. Sun, Enhanced visible light photocatalytic degradation of methylene blue by F-doped TiO_2, Appl. Surf. Sci. 319 (2014) 107-112. https://doi.org/10.1016/j.apsusc.2014.07.038

[22] R. P. Barkul, M. K. Patil, S. M. Patil, V. B. Shevale, S. D. Delekar, Sunlight-assisted photocatalytic degradation of textile effluent and Rhodamine B by using iodine doped TiO_2 nanoparticles, J. Photoch. Photobio. A. Chem. 349 (2017) 138-147. https://doi.org/10.1016/j.jphotochem.2017.09.011

[23] M. Sathish, B. Viswanathan, R. P. Viswanath, C. S. Gopinath, Synthesis, Characterization, Electronic Structure, and Photocatalytic Activity of Nitrogen-Doped TiO_2 Nanocatalyst, Chem. Mater. 17 (25) (2005) 6349-6353. https://doi.org/10.1021/cm052047v

[24] S. Demircia, T. Dikici, M. Yurddaskal, S. Gultekind, M. Toparli, E. Celik, Synthesis and characterization of Ag doped TiO_2 heterojunction films and their photocatalytic performances, Appl. Surf. Sci. 390 (2016) 591-601. https://doi.org/10.1016/j.apsusc.2016.08.145

[25] S. Murcia-López, M. C. Hidalgo, J. A. Navío, Synthesis, characterization and photocatalytic activity of Bi-doped TiO_2 photocatalysts under simulated solar irradiation, Appl. Catal. A . Gen. 404 (2011) 404, 59-67. https://doi.org/10.1016/j.apcata.2011.07.008

[26] L. Elsellami, H. Lachheb, A. Houas, Synthesis, characterization and photocatalytic activity of Li-, Cd-, and La-doped TiO_2, Mat. Sci. Semicon. Proc. 36 (2015) 103-114. https://doi.org/10.1016/j.mssp.2015.03.032

[27] N. Aman, P. K. Satapathy, T. Mishraa, M. Mahatoa, N. N. Das, Synthesis and photocatalytic activity of mesoporous cerium doped TiO_2 as visible light sensitive photocatalyst, Mater. Res. Bull. 47 (2012)179-183. https://doi.org/10.1016/j.materresbull.2011.11.049

[28] Y. L. Kuoa, T. L. Su, F. C. Kung, T. J. Wu, A study of parameter setting and characterization of visible-light driven nitrogen-modified commercial TiO_2 photocatalysts, J. Hazard. Mater. 190 (2011) 938-944. https://doi.org/10.1016/j.jhazmat.2011.04.031

[29] S. Liu, X. Chen, A visible light response TiO_2 photocatalyst realized by cationic S-doping and its application for phenol degradation, J. Hazard. Mater. 152 (1) (2008) 48-55. https://doi.org/10.1016/j.jhazmat.2007.06.062

[30] Y. Liu, Z. Wang, W. Fan, Z. Geng, L. Feng, Enhancement of the photocatalytic performance of Ni-loaded TiO_2 photocatalyst under sunlight, Ceram. Int. 40 (3) (2014) 3887-3893. https://doi.org/10.1016/j.ceramint.2013.08.030

[31] J. Zhou, Y. Zhang, X. S. Zhao, A. K. Ray, Photodegradation of Benzoic Acid over Metal-Doped TiO_2, Ind. Eng. Chem. Res. 45 (10) (2006) 3503-3511. https://doi.org/10.1021/ie051098z

[32] P. Margan, M. Haghighi, Hydrothermal-assisted sol–gel synthesis of Cd-doped TiO_2 nanophotocatalyst for removal of acid orange from wastewater, J. Sol-Gel. Sci. Tech. 81 (2017) 556-569. https://doi.org/10.1007/s10971-016-4217-7

[33] S. F. Chen, Y. Z. Liu, Study on the photocatalytic degradation of glyphosate by TiO2 photocatalyst, Chemosphere. 67 (5) (2007) 1010-1017. https://doi.org/10.1016/j.chemosphere.2006.10.054

[34] M. R. Hoffmann, S. T. Martin, W. Choi, D. W. Bahnemann, Environmental Applications of Semiconductor Photocatalysis, Chem. Rev. 95(1) (1995) 69-96. https://doi.org/10.1021/cr00033a004

[35] C. Sahoo, A. K. Gupta, I. M. S. Pillai, Photocatalytic degradation of methylene blue dye from aqueous solution using silver ion-doped TiO_2 and its application to the degradation of real textile wastewater, J. Environ. Sci. Heal. A. 47 (2012) 1428-1438. https://doi.org/10.1080/10934529.2012.672387

[36] C. H. Chiou, R. S. Juang, Photocatalytic degradation of phenol in aqueous solutions by Pr-doped TiO_2 nanoparticles, J. Hazard. Mater. 149 (2007) 1-7. https://doi.org/10.1016/j.jhazmat.2007.03.035

[37] A. W. Xu, Y. Gao, H. Q. Liu, The preparation, characterization, and their photocatalytic activities of rare-earth-doped TiO_2 nanoparticles, J. Catal. 207 (2002) 151-157. https://doi.org/10.1006/jcat.2002.3539

[38] S. Chen, W. Zhao, W. Liu, H. Zhang, X. Yu, Preparation, characterization and activity evaluation of p–n junction photocatalyst p-$CaFe_2O_4$/ n-ZnO, Chem. Eng. J. 155 (2009) 466-473. https://doi.org/10.1016/j.cej.2009.07.009

[39] J. Shi, J. Zheng, Y. Hu, Y. Zhao, Photocatalytic Degradation of Methyl Orange in Water by Samarium-Doped TiO_2, Environ. Eng. Sci. 25 (4) (2008) 489-496. https://doi.org/10.1089/ees.2007.0048

[40] S. Sohrabnezhad, A. Pourahmad, E. Radaee, Photocatalytic degradation of basic blue 9 by CoS nanoparticles supported on AlMCM-41 material as a catalyst, J. Hazard. Mater. 170 (2009) 184-190. https://doi.org/10.1016/j.jhazmat.2009.04.108

[41] K. M. Reza, A. Kurny, F. Gulshan, Parameters affecting the photocatalytic degradation of dyes using TiO_2: A review, Appl. Water. Sci. 7 (2017) 1569-1578. https://doi.org/10.1007/s13201-015-0367-y

[42] C. Sahoo, A. K. Gupta, A. Pal, Photocatalytic degradation of Methyl Red dye in aqueous solutions under UV irradiation using Ag^+ doped TiO_2, Desalination. 181 (2005) 91-100. https://doi.org/10.1016/j.desal.2005.02.014

[43] M. Hamadanian, A. Reisi-Vanani, A. Majedi, Preparation and characterization of S-doped TiO_2 nanoparticles, effect of calcination temperature and evaluation of photocatalytic activity, Mater. Chem. Phys. 116 (2009) 376-382. https://doi.org/10.1016/j.matchemphys.2009.03.039

[44] J. A. Cortes, M. T. Alarcon-Herrera, M. Villican -Mendez, J. Gonzalez-Hernandez, J. F. Perez-Robles, Impact of the Kind of Ultraviolet Light on the Photocatalytic Degradation Kinetics of the TiO_2/UV Process, Environ. Prog. Sustain. 30 (2011) 318-325. https://doi.org/10.1002/ep.10480

[45] P. Sathishkumar, R.V. Mangalaraja, S. Anandan, M. Ashokkumar, $CoFe_2O_4/TiO_2$ nanocatalysts for the photocatalytic degradation of Reactive Red 120 in aqueous solutions in the presence and absence of electron acceptors, Chem. Eng. J. 220 (2013) 302-310. https://doi.org/10.1016/j.cej.2013.01.036

[46] S. Sakthivel, B. Neppolian, M.V. Shankar, B. Arabindoo, M. Palanichamy, V. Murugesan, Solar photocatalytic degradation of azo dye: comparison of photocatalytic efficiency of ZnO and TiO_2, Sol. Energ. Mat. Sol. C. 77 (2003) 65-82. https://doi.org/10.1016/S0927-0248(02)00255-6

[47] N. Kumar, H. Mittal, S. M. Alhassan, S. S. Ray, Bionanocomposite hydrogel for the adsorption of dye and reusability of generated waste for the photo-degradation of ciprofloxacin: A demonstration of the circularity concept for water purification, ACS. Sustain. Chem. Eng. 6 (12) (2018) 17011-17025. https://doi.org/10.1021/acssuschemeng.8b04347

[48] F. Li, S. Sun, Y. Jiang, M. Xia, M. Sun, B. Xue, Photodegradation of an azo dye using immobilized nanoparticles of TiO_2 supported by natural porous mineral, J. Hazard. Mater. 152 (2008) 1037-1044. https://doi.org/10.1016/j.jhazmat.2007.07.114

[49] J. Li, L. Li, L. Zheng, Y. Xian, L. Jin, Photoelectro- catalytic degradation of rhodamine B using Ti/TiO_2 electrode prepared by laser calcination method, Electrochimica Acta. 51 (2006) 4942-4949. https://doi.org/10.1016/j.electacta.2006.01.037

[50] A. R. Khataee, M. B. Kasiri, Photocatalytic degradation of organic dyes in the presence of nanostructured titanium dioxide: Influence of the chemical structure of dyes, J. Mol. Catal. A: Chem. 328 (2010) 8-26. https://doi.org/10.1016/j.molcata.2010.05.023

[51] N. Barka, S. Qourzal, A. Assabbane, A. Nounah, Y. Ait-Ichou, Factors influencing the photocatalytic degradation of Rhodamine B by TiO_2-coated non-woven paper, J. Photoch. Photobio. A. 195 (2008)346-351. https://doi.org/10.1016/j.jphotochem.2007.10.022

[52] H. Khan, I. K. Swati, Fe^{3+}-doped Anatase TiO_2 with d–d Transition, Oxygen Vacancies and Ti^{3+} Centers: Synthesis, Characterization, UV–vis Photocatalytic and Mechanistic Studies, Ind. Eng. Chem. Res. 55 (2016) 6619-6633. https://doi.org/10.1021/acs.iecr.6b01104

[53] T. Ali, P. Tripathi, A. Azam, W. Raza, A. S. Ahmed, A. Ahmed, M. Muneer, Photocatalytic performance of Fe-doped TiO_2 nanoparticles under visible-light irradiation, Mater. Res. Express. 4 (2017) 015022.

[54] P. Margan, M. Haghighi, Hydrothermal-assisted sol–gel synthesis of Cd-doped TiO_2 nanophotocatalyst for removal of acid orange from wastewater, J. Sol-gel. Sci. Techn. 81 (2017) 556-569. https://doi.org/10.1007/s10971-016-4217-7

[55] C. Sahoo, A. K. Gupta, I. M. S. Pillai, Photocatalytic degradation of methylene blue dye from aqueous solution using silver ion-doped TiO_2 and its application to the degradation of real textile wastewater, J. Environ. Sci. Heal. A. 47 (10) (2012) 1428-1438. https://doi.org/10.1080/10934529.2012.672387

[56] Q. Xiao, J. Zhang, C. Xiao, Z. Si, X. Tan, Solar photocatalytic degradation of methylene blue in carbon-doped TiO_2 nanoparticles suspension, Sol. Energy. 82 (8) (2008) 706-713. https://doi.org/10.1016/j.solener.2008.02.006

[57] M. Shaban, A. M. Ahmed, N. Shehata, M. A. Betiha, A. M. Rabie, Ni-doped and Ni/Cr co-doped TiO_2 nanotubes for enhancement of photocatalytic degradation of

methylene blue, J. Colloid. Interf. Sci. 555 (1) (2019) 31-41.

https://doi.org/10.1016/j.jcis.2019.07.070