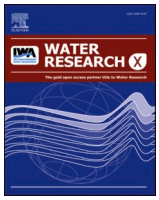
Water Research X 22 (2024) 100217

Contents lists available at ScienceDirect

Water Research X

journal homepage: www.sciencedirect.com/journal/water-research-x



# Efficient smartphone-based measurement of phosphorus in water

Haiping Ai, Kai Zhang, Huichun Zhang *

*Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, OH, 44106, USA*

ARTICLE INFO

*Keywords:*
Colorimetric analysis
Machine learning
Phosphorus monitoring
Smartphone

ABSTRACT

Agricultural runoff is one of the main sources of excess phosphorus (P) in different water bodies, subsequently leading to eutrophication and harmful algal blooms. To effectively monitor P levels in water, there is a need for simple measurement tools and extensive public involvement to enable regular and widespread sampling. Several smartphone-based P measurement methods have been reported, which extract red-green-blue (RGB) values from colorimetric reactions to build statistical regression models for P quantification. However, these methods typically require meticulous light conditions, involve initial equipment investment, and have undergone limited testing for large-scale applications. To overcome these limitations, this study developed a smartphone-based, equipment-free and facile P colorimetric analysis method. Following the standard procedure of the ascorbic acid approach, colorimetric reactions were captured by a smartphone camera, and RGB values were extracted using Python code for modeling. Different indoor light conditions, phone types, containers, and types of water samples were examined, resulting in a collection of 1922 images. The best regression model, employing random forest with RGB values and container types as inputs, achieved an $R^2$ of 0.97 and an RMSE of 0.051 for P concentrations ranging from 0.01 to 1.0 mg P/L. Additionally, the optimal classification model could estimate the level of P below 0.1 mg P/L with an accuracy of 95.2 (or 77.4 % for <0.05 mg P/L). The strong performance of the developed models, which are also available freely online, suggests an easy and effective tool for more frequent P measurement and greater public involvement.

## Introduction

Excess nutrients in surface waters worldwide cause eutrophication, leading to excessive growth of autotrophs like algae and cyanobacteria (Khan et al., 2014). Apart from wastewater discharge, agriculture and animal husbandry are two main nutrient sources leading to eutrophication (Sala and Mujeriego, 2001). In recent decades, Lake Erie has been plagued by harmful algal blooms, with soluble reactive phosphorus, primarily originating from non-point agricultural runoffs, identified as one of the key driving factors (Ai et al., 2023; Ho and Michalak, 2017). Effective P monitoring and control are critical for mitigating eutrophication and harmful algal blooms. Capturing transient changes in P concentration over time is another critical step, as the concentration of P in runoffs is significantly influenced by variations in weather and fertilizer application. For example, sudden precipitation following the application of fertilizer can result in high P levels in runoffs, while prolonged precipitation can interrupt fertilizer application and reduce P in agricultural runoffs (Guo et al., 2020; Michalak et al., 2013). Moreover, smaller water bodies, such as ponds and small lakes, are often overlooked in environmental monitoring due to their sheer number and small sizes. However, their pollution poses significant public health concerns because of their proximity to communities (Downing, 2010; Mrdjen et al., 2018). Regular monitoring of ditches, creeks, and lakes that captures the spatial and temporal changes of P can not only help identify hot spots but also support more targeted nutrient management plans. In this context, an effective and easy method for measuring P is in great demand.

Several methods exist for measuring P, with new techniques continually emerging; however, most of current methods have certain limitations, particularly for widespread monitoring by the general public or farm owners. For instance, ion chromatography (IC) can efficiently analyze low P concentrations ranging from μg/L to 1 mg/L (Scientific, 2023). However, IC is expensive, delicate and vulnerable to background anions. Another common method is the ascorbic acid method (EPA, 1978; Murphy and Riley, 1962), which involves colorimetric reactions of P and the analysis by a spectrometer. However, the spectrometer not only is expensive but also requires a well-controlled lab environment. Although size-reduced spectrometers are available at

* Corresponding author.
*E-mail address:* hjz13@case.edu (H. Zhang).

https://doi.org/10.1016/j.wroa.2024.100217
Received 16 January 2024; Received in revised form 3 March 2024; Accepted 4 March 2024
Available online 5 March 2024

a cost of a few hundred dollars (e.g., $310 on Amazon), their widespread distribution remains infeasible and their detection limits are often not specified. Therefore, a simple and affordable P measurement method is needed to enable wide application.

The ubiquitous smartphone and its increasing capability provide a promising tool for public involvement in citizen science (Zheng et al., 2022). Simplicity and effectiveness are crucial for citizen science applications. Without any advanced instruments, a few studies (Table 1) have employed smartphone cameras to take pictures of colorimetric reactions, and then extracted the RGB values of the images to estimate the P concentrations. However, those methods require professional or customized auxiliaries, such as LED-equipped boxes, synthesized test strips, expensive cuvettes, or complex fluorescent methods, leaving considerable challenges for their applications in public involvement. The need for any investment in equipment could deter widespread public engagement. Only one study (Table 1) directly used a smartphone for P measurement without any additional equipment (Costa et al., 2020). However, the reported performance was compromised, showing a relative error (RE=$\frac{|\text{residuals}|}{\text{actual concentration}} \times$ 100%) of 70 % for detection ranging from 0.25 to 10 mg P/L. For P measurement without any additional instruments, employing a machine-learning (ML) model could compensate for performance deterioration or enable a rapid concentration range determination, referring to a drinking water chlorine residual estimation study that achieved an accuracy of 94 % in a 2-level classification by a random forest (RF) model (Schubert et al., 2022).

Additionally, the previously developed phone-based P measurement methods (as shown in Table 1) paid insufficient attention to quality control. For example, the validity of calibration curves, usually based on a small number of data points (mostly 5–10), is not well-established, as the errors introduced by device setup and calibrations in the real application are often not considered, and the reproducibility and robustness of measurements are rarely examined. In addition, the discussions often lack details on measurement deviations, and errors are obscured by mean values. There is also a noticeable lack of results and discussion regarding low concentration (0–0.1 mg P/L) P detection, even though regulatory guidelines suggest limits of 0.05 mg P/L in streams entering lakes and 0.1 mg P/L in running waters (Litke, 1999).

In this study, we developed an easy and effective P measurement method with the smartphone camera coupled with an ML algorithm. With the aid of an ML tool, we significantly simplified the measurement process by eliminating the need for any equipment used in the reported phone-based methods. In total, 1922 images of colorimetric reaction were collected under different indoor light conditions, phone types, containers, and water sample types. Moreover, we focused on P concentrations ranging from 0.01 to 1.0 mg P/L, particularly studying real water samples. Subsequently, RGB values were extracted from the colorimetric images to build accurate regression and classification ML models with their corresponding concentrations. In summary, this study developed an easy, versatile, and effective P measurement method using smartphones, requiring minimal equipment and suitable for public use.

## Results and discussion

### *Development of ML regression models*

As the model performance shown in Table S1, except for convolutional neural network (CNN) and multilinear regression (MLR), all five tree-based ML regression models achieved comparable performance on the test set for P measurement. In this study, we chose to use RF as it can evaluate the input feature importance by the permutation approach (Ai et al., 2023), which can help evaluate the feature importance below. As for the permutation approach, it calculates the feature importance by shuffling the order an input feature values and comparing the model performance before and after the shuffling (Jones and Linder, 2015).

CNN is highly effective at tasks such as image classification, object detection, and image segmentation (Girshick et al., 2015). However, with a poor $R^2$ value of −0.03, it did not effectively discern the intensity changes of the colorimetric measurements. The MLR model showed a strong correlation between P concentrations and RGB values with an $R^2$ of 0.82. However, this correlation is not sufficient for accurate P estimation when compared to the ML models developed in this study.

### *Determination of P measurement conditions*

Unlike sophisticated instrument analysis, an easy and portable method that requires minimal training, effort, and no capital investment for the general public can justify a sacrifice in accuracy (Li et al., 2022). As the image quality was likely affected by operational parameters, we examined different phones, indoor light conditions, and container types

**Table 1**
Summary of smartphone-based phosphate measurement studies since 2015. RSD = relative standard deviation.

| Conc. range, mg P/L | LOD mg P/L | Colorimetric method | Image object | Device for imaging | Performance | Reference |
|---|---|---|---|---|---|---|
| 0.3–1.0 | 0.01 | Ascorbic acid method | Test tube | LED light box | $R^2$=0.993; RSD=0.95 % and 2.1 % for 0.5 and 1.0 mg P/L, respectively. | (Moonrungsee et al., 2015) |
| 0.1–5.0 | 0.016 | Ascorbic acid method | Cuvette | LED tool kit | $R^2$=0.99; mean RSD=0.95 % for 0.5, 1.5, 3, and 5.0 mg P/L. | (Das et al., 2022) |
| 0.056–7.75 | 0.0167 | Ascorbic acid method | Cuvette | LED light box | $R^2$=0.999; mean RSD=3.52 % for 0.056–0.93 mg P/L; mean RSD=5.3 % for 0.338–7.75 mg P/L. | (Li et al., 2022) |
| 0.3–28.0 | 0.001 | Ascorbic acid method | Cuvette | LED light box | $R^2$=0.96 for 0–1.0 mg P/L; mean RSD=2.01 % for 0.3–28 mg P/L. | (Lavanya et al., 2023) |
| 0.25–10.0 | 1.6 | Ascorbic acid method | Cuvette | LED light box | $R^2$=0.980; RSD varied from 2.3 % to 9.8 % for 5 replicates of 1.0 mg P/L. | (de Souza et al., 2023) |
| 0.1–10 | 0.028 | Ascorbic acid method | Plate | LED light box | $R^2$=0.978; RSD = 1.34 % for 6 replicates of 5.0 mg P/L. | (Xing et al., 2022) |
| 0.25–10.0 | NA | Commercial liquid test kit | Glass bottle | None | Relative error is about 30 % and 70 % for corrected and uncorrected measurements, respectively. | (Costa et al., 2020) |
| 0.155–1.24 | 0.031 | Fluorometric heptamolybdate method | Cuvette | LED tool kit | $R^2$=0.989; mean RSD=1.2 %. | (Granica and Tymecki, 2019) |
| 0.0155–1.55 | 0.0037 | Lanthanide fluorescent method | Cuvette | UV light box | $R^2$=0.996 | (Wu et al., 2022) |
| 0.062–1.98 | 0.031 | Wax-printed paper strip | Paper strip | UV light box | $R^2$=0.96; mean RSD=12.67 %. | (Sarwar et al., 2019) |
| 0.2–8.3 | NA | TMB oxidation method | Designed microplate | UV light box | $R^2$=0.985 | (Li et al., 2020) |
| 0.031–23.25 | 0.093 | Commercial test strips | Paper strip | Infrared lightbox, camera, processer | $R^2$=0.998 | (Heidari-Bafroui et al., 2021) |

to capture the colorimetric differences in P solutions. During the RF model development, we started with all input features for the first RF model, then eliminated the indoor light, phone, and container features sequentially to build the second, third, and fourth RF model based on the feature importance. The model performance and feature importance are summarized in Table 2. The first RF model achieved an $R^2$ of 0.98 and an RMSE of 0.049. Given the model's impressive performance, it was necessary to evaluate the importance of the input features to identify and exclude any unnecessary ones for a simpler method. In the first RF model with all the input features, the light feature showed the lowest importance (0.002) and was thus excluded in the second RF model.

Excluding the light feature in the second RF model maintained surprisingly strong performance, yielding an $R^2$ of 0.97 and an RMSE of 0.051, on par with the first RF model. However, almost all reported studies used light boxes to control the light condition and used LED as the light source (Table 1). This approach ensures consistent RGB values for images of designated P concentrations (Li et al., 2022). Indeed, we observed a significant RGB variation due to varying light conditions among images from samples with the same concentration (Figs. S1–2). This also explained the poor performance of the MLR model (Table S1) as it did not consider indoor light condition and container type in the model input. However, when the data groups under bright, dim, and dark conditions in the second RF model were compared, P estimation performed similarly under all three conditions (Fig. 1a). This is likely because there is a clear difference in color intensity under the bright, dim, and dark conditions, as indicated by the plots of mean intensity against P concentration and the different mean intensity ranges for bright, dim, and dark conditions (Fig. S2, ANOVA, $p < 0.001$). Giving the large amount of data used to train the ML model, the model has likely already learned the differences among the three indoor light conditions. When the model encounters a query sample, it could 'correctly' determine the indoor light condition as bright, dim or dark based on the observed light intensity against the respective mean light intensity, even without explicit input. Therefore, different indoor light conditions have no significant influence on the P estimation using our model. In comparison, the reported statistical models have only a small number of standards. Therefore, using light boxes to fix the light condition becomes necessary to reproduce the standard values for statistical models.

As consistent RGB values in previous studies, achieved through meticulous control of light conditions, including box dimensions, the number of LED lights, and power supplies, are essential for simple statistical models (studies in Table 1), these results demonstrate that our model can circumvent the need for equipment to control light, providing a significant simplification for smartphone-based P measurement.

In the third RF model, the phone feature was excluded as it had the lowest feature importance (0.006) in the second RF model (Table 2). The third RF model still maintained good performance with an $R^2$ of 0.97 and an RMSE of 0.051. In this model, data in the iOS and Android sets showed similar performance in P estimation (Fig. 1b). Both phones used in the study had high camera resolutions of 48 and 40 megapixels, respectively, which is essential for effective image capturing (Schubert et al., 2022). It can be inferred that different phones with high camera resolutions should function similarly in capturing colorimetric images. Therefore, it is not necessary to include the 'phone' feature in the model input.

Lastly, the container feature was excluded in the fourth RF model. Notably, the model's performance deteriorated, with an $R^2$ of 0.93 and an RMSE of 0.082 (Table 2). Therefore, the fourth RF model, which relies only on RGB values, is not capable of making effective P estimations, although it is still much better than the MLR model. When comparisons were made among the datasets using a cuvette, a bottle, and a dish as the container (Fig. 1c), dish showed the best performance, followed by bottle and cuvette. This discrepancy is reasonable considering that the three container types have different surfaces for color capturing. As shown in Materials and Methods, colorimetric solutions can be directly captured by a camera above the liquid surface in dishes. In contrast, the cylindrical shape of glass bottles might introduce reflection into the colorimetric solution (Costa et al., 2020), so do cuvette walls. In practical terms, low-cost white plastic dish-like containers (common sauce dishes) are readily available and well-suited for P measurement. Referring to Table 1, most studies utilized more expensive containers, such as cuvettes, for testing. Based on the above results, we will focus on the third RF model which includes both container and RGB features in the input.

*Comparison with similar reported methods*

The first six studies in Table 1 employed LED light boxes for P measurement based on the ascorbic acid method. As a reference, these studies achieved $R^2$ values ranging from 0.96 to 0.9988 and RSD ranging from 0.95 % to 9.8 %. The performance of these methods was determined based on linear, log-scaled linear, or exponential data fitting between mean concentrations and RGB values, where 1 to 3 color channels were used. Compared with the limited data fitting at high P concentration ranges (0.056–28 mg P/L) in those studies, our RF model achieved an $R^2$ of 0.995, an RMSE of 0.021 (Fig. 2a), and an RSD of 6.07 % for the concentration range of 0.01–1.0 mg P/L, with a large number (1922) of data points. However, there are three limitations in those statistical models. Firstly, the applicability of those models is not fully evaluated. They were developed based on a small number of data points (5–20 mean values) across a concentration range of 0.0558 to 30 mg P/L. Mean values can obscure data deviation in the fitting process, potentially decreasing the RSD. Additionally, the P concentrations studied were mostly higher than 0.1 mg P/L, where higher concentrations can mask relative measurement errors/standard deviations. Secondly, as mentioned earlier, the statistical models require consistent RGB-concentration values to build a reliable relationship, so those studies required data calibration before measurement, which would significantly increase the workload for the measurement of real water samples. Thirdly, with comparable performance, our RF model eliminates the need for any equipment. The RF model collects over 1900

**Table 2**
Overall model performance and feature importance in four RF models during the feature exclusion process.

| RF model | Performance | | Feature importance | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMSE | R | G | B | Container | Phone | Light |
| First | 0.98 (0.004) | 0.049 (0.004) | 0.778 (0.006) | 0.075 (0.006) | 0.086 (0.003) | 0.054 (0.002) | 0.006 (0.001) | 0.002 (0) |
| Second | 0.97 (0.004) | 0.051 (0.004) | 0.778 (0.006) | 0.075 (0.006) | 0.087 (0.003) | 0.054 (0.002) | 0.006 (0.001) | – |
| **Third** | **0.97 (0.004)** | **0.051 (0.004)** | 0.779 (0.006) | 0.077 (0.006) | 0.090 (0.002) | 0.055 (0.002) | – | – |
| Fourth | 0.93 (0.006) | 0.082 (0.004) | 0.800 (0.006) | 0.088 (0.005) | 0.111 (0.003) | – | – | – |

Note: The dataset was randomly stratified split to training and test sets in a 4:1 ratio for 20 times by changing the random state from 42 to 61. The model performance against the test sets and the feature importance values are presented as mean values, and the value in parenthesis is the standard deviation.


**Fig. 1.** Model performance for each data group based on different (a) indoor light conditions, (b) phone inputs (IOS and Android), and (c) container types in the second, third, and fourth RF model, respectively. Note: The data dataset was randomly stratified split to training and test set in a 4:1 ratio for 20 times by changing the random state from 42 to 61. The model performance is presented as the mean value, and the error bar is the standard deviation.

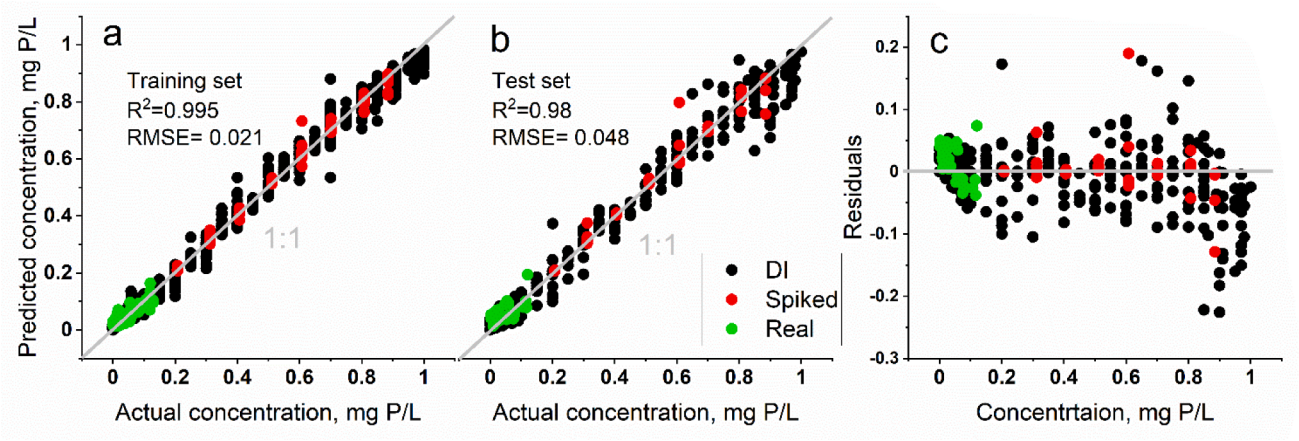



**Fig. 2.** Scatter plot of predicted and actual P concentrations in the (a) training set and (b) test set of the third RF model. The gray line is the reference of the 1:1 slope; (c) Prediction residuals for the test set (Prediction residuals = predicted value – actual value). Based on the water sample sources, the test set data was split into DI (pure water-made P solution, black), spiked (real water-made P solution, red), and real (in-situ samples, green). **Note**: the model performance was based on a random state of 42 as the representative results as the standard deviation of $R^2$ and RMSE is negligible for 20 times of data split (Table 2).

images across various conditions and sampling environments, representing a truly equipment-free method for citizen science and wide application, potentially broadening the end-user.

Based on the third RF model, predicted and actual P concentrations in the test set are plotted in Fig. 2b. As specified in Materials and Methods, the colorimetric P solutions were composed of DI water, spiked real water, and real water. These different water matrices were used to examine the method's generalization ability beyond laboratory conditions. The predicted data from these three sets followed a similar pattern along the 1:1 reference line (Fig. 2b). Furthermore, the prediction residuals for the spiked and real groups fell within the range of the DI water group (Fig. 2c). The limit of detection (LOD) was determined to be 0.036 mg P/L. Fig. 2c also indicates that there is slight overestimation for low concentrations and underestimation for high concentrations, but no significant prediction bias for mid-range concentrations, except for a few outliers. Note that the spike concentrations were between 0.2 and 0.9 mg P/L, resulting in a mean RE of 3.52 %. Compared with the only reported smartphone-method without a light box (Costa et al., 2020) (Table 1), which achieved a RE of 70 % (additional empirical correction can reduce to 30 %) in the range of 0.25 to 10 mg P/L, our study's RF model performed much better, with an RE of 8.43 % for the test set within the concentration range of 0.1 to 1.0 mg P/L, further highlighting the robustness of our method.

Apart from the commonly used ascorbic acid method, other studies, like the 8th to 10th in Table 1, have employed fluorescent methods to increase measurement sensitivity and specificity. However, when employing smartphones, these methods still rely on extracting RGB values from images. This extraction is based on the interaction between phosphate and heptamolybdate, lanthanide, or phosphate-binding protein, and an additional light source is necessary for capturing the images. In terms of performance, these methods perform similar to those based on the ascorbic acid method, with the $R^2$ values ranging from 0.96 to 0.9962 and RSD ranging from 1.2 % to 12.67 % across concentration ranges of 0.0155 to 1.98 mg P/L, using the same data fitting methods. Considering the complexity of operations, stringent measurement requirements, and initial hardware investment of these methods, our model still demonstrates its superiority.

### *Classification models for real water samples*

As mentioned in Introduction, the suggested total P limit is 0.1 mg P/L for running water and 0.05 mg P/L for water entering lakes (Litke,

1999). The LOD and RMSE for the RF regression model were 0.036 mg P/L and 0.048, respectively. However, the performance of the RF regression model on real water samples only showed an $R^2$ of 0.098, an RMSE of 0.031, and an RSD of 81.56 % for the concentration range of 0–0.129 mg P/L (Fig. 2b). Therefore, the regression model may not be capable of accurately quantifying low P concentrations. To this end, we developed a classification model to rapidly classify P levels in water samples into four levels: L1 (0–0.05/0.1 mg P/L), L2 (0.05/0.1–0.4 mg P/L), L3 (0.4–0.7 mg P/L), and L4 (0.7–1.0 mg P/L), as detailed in Table S2 and further explained in the Materials and Method. This model assists in identifying whether P levels exceed limits or indicate the extent of eutrophication, benefiting the scientific community and decision-makers. As shown in Fig. S3, both classification models demonstrate high accuracy. When L1 is set at <0.05 mg P/L, the overall accuracy is 86.6 % with a Kappa coefficient of 0.82; when L1 is set at <0.1 mg P/L, the overall accuracy is 91.1 % with a Kappa coefficient of 0.88. Note that a Kappa value of >0.8 indicates strong agreement (McHugh, 2012). Additionally, when the RF classification model was applied to the second in-situ sampling data, covering low P concentrations (0.001–0.129 mg P/L, encompassing L1 and L2, see Fig. S4, Table S3, and Materials and Methods), the overall accuracy was 62.7 % (L1=57.9 %, L2=68.9 %, 0.05-limit) and 88.2 % (L1=97.7 %, L2=33.3 %, 0.1-limit) for the two models. Note that in the above models, we employed all previous data, including the first in-situ sampling data, for the training set and the second in-situ sampling data for the test set. The low L2 accuracy (33.3 %) for the 0.1-limit might be due to the low P concentrations of the real water samples. Only two sites of samples (sewage effluent and farmland ditch) had P concentrations slightly higher than 0.1 mg P/L (0.129 mg P/L).

## Conclusion

This study developed an accessible and facile P measurement method using smartphone cameras and RF models. The RF regression model achieved an $R^2$ of 0.97 and an RMSE of 0.051 for the concentration range of 0.01–1.0 mg P/L with the LOD of 0.036 mg P/L. Furthermore, overcoming limitations of traditional methods, this work offers reliable P concentration classifications in the critical range of 0 to 0.1 mg P/L, where the regression model is incapable of quantifying such low P concentrations, making the classification models pivotal for monitoring P in water bodies. The RF models, designed to be user-friendly and to minimize equipment needs, are available online for public use (https://envmodel-cwru.streamlit.app/), promoting widespread participation in environmental monitoring. Field tests in the great Cleveland area demonstrated the method's real-world applicability, even though some measurement errors were noted at lower concentrations (0–0.13 mg P/L).

The ascorbic acid method requires solution preparation and sample handling before smartphone application, which may deter some uses from the public. This may be especially true when pretreatment is needed for samples with interfering substances and color. Despite that, compared with previous studies (Table 1), our approach significantly simplifies measurement requirements. Reagents for the ascorbic acid method can be pre-packaged according to the specified ratios, like a covid test kit, and distributed to people closely involved in nutrient management, such as farm owners, park rangers, and community volunteers. These reagent solutions can be stored for weeks before mixing, and the mixed solution remains stable for 24 h. Overall, this new approach not only highlights the potential of machine learning in environmental applications but also paves the way for greater public engagement in water quality management, thereby enhancing awareness and improving responses to P pollution challenges.

## Materials and methods

### *Ascorbic acid method*

The ascorbic acid method (Murphy and Riley, 1962) used here is the most commonly used colorimetric method for soluble phosphorus determination. Generally, the method requires mixing 5 N sulfuric acid, 2.743 g/L potassium antimonyl tartrate solution, 40 g/L ammonium molybdate solution, and 17.6 g/L ascorbic acid solution in a volume ratio of 5:0.5:1.5:3. During the colorimetric reaction, phosphate, ammonium molybdate, and potassium antimony tartrate form a phosphomolybdate complex in acidic conditions. This complex is then reduced by ascorbic acid to generate a phosphor-antimonyl molybdenum blue substance. The color becomes stable after 10 min, and the intensity of the blue color is linearly proportional to the P concentration. The detection range of the method is 0–1.2 mg P/L. In addition, arsenate and high concentrations of iron may interfere with the measurement. This interference occurs because the structure of arsenate is highly similar to that of phosphate, and iron can react with the reducing reagents in colorimetric tests. The stock P solution is prepared to be 10 mg P/L using $KH_2PO_4$, with $KH_2PO_4$ powder dried at 105 °C for an hour to eliminate moisture before being used for solution preparation. Unless specified otherwise, all mentioned P concentrations refer to elemental P. All chemicals were purchased from Fisher Scientific, and all solutions were prepared using Millipore nanopure water (DI).

For the P measurement, the P stock solution was diluted to concentrations ranging from 0.01 to 0.1 mg/L at intervals of 0.01 mg/L, and from 0.1 to 1.0 mg/L at intervals of 0.05 mg/L. 4 mL of mixing reagent solution and 25 mL of the diluted P solution were mixed in a volume ratio of 4:25 (at the same ratio, the volume can be adjusted per the application) and sonicated for 5 s for better mixing. After 10 min of reaction, the blue color stabilized. The reaction solution was then transferred to a 1-cm cuvette and analyzed using a UV/Vis spectrometer (Cary 60, Agilent Technologies, USA) at a wavelength of 880 nm within 1 h. Absorbance values obtained from the spectrometer were plotted against known P concentrations to determine the calibration curves. In the ascorbic acid method published by the USEPA (EPA, 1978), the detection range is 0.3–1.2 mg P/L for a 1-cm cuvette and 0.01–0.3 mg P/L for a 5-cm cuvette. In this study, we used 1-cm cuvette for concentrations ranging from 0.01 to 1 mg P/L but still obtained excellent and consistent calibration relationships for both the full concentration (0–1.0 mg/L) and low concentration (0–0.1 mg P/L) ranges, with similar slopes (0.644 vs. 0.646) and $R^2$ values (1.0 vs. 0.99), as shown in Fig. S5.

### *P colorimetric image analysis*

To examine the versatility of the method, we tested phones with two different operating systems (android and IOS), three light conditions, three types of containers, and three types of water samples to generate 14 groups of data, as shown in Fig. 3 and Table 3. After the above P measurement by a spectrometer, the cuvette was placed against a white paper background for image capture using an iPhone 14 Pro (iOS system, Apple, rear camera resolution of 48 megapixels) and a Mate 20 Pro (Android system, Huawei, rear camera resolution of 40 megapixels). The distance between the camera and the cuvette was approximately 10–20 cm, and each sample was centered and photographed three times. Additionally, three different indoor light conditions (lab room light on [bright], lab room light off [dark], and hallway [dim]), and two other containers (a 40-mL glass bottle and a 5-cm wide white plastic dish, same image taking method as cuvette) were similarly examined to assess the influence of light and containers on the performance of machine learning models. All images were collected across 14 experimental events on different dates and times, further examining the method's generalization capability for varying surrounding environments. Based on the Python code, an image center area of 50 × 50 pixels was selected and averaged for RGB values extraction to reduce the variation of the

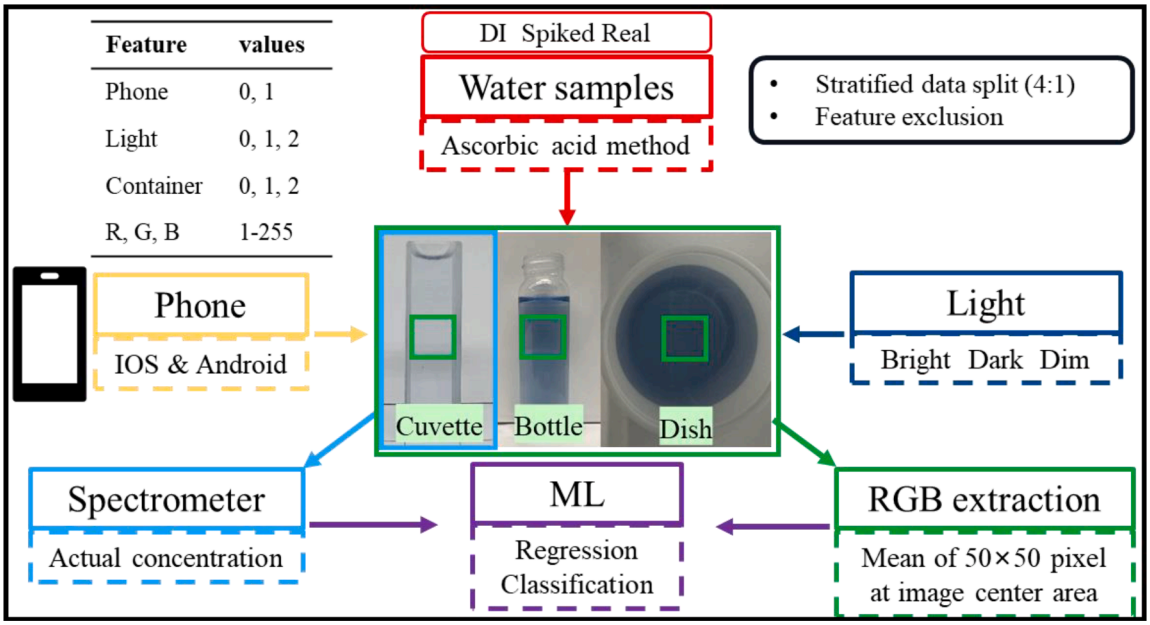



**Fig. 3.** Experimental setup and the workflow.

**Table 3**
Summary of labels, data conditions, and amount of data for each condition.

| Label | Phone | Light | Container | Water type | Amount of data | Conc. mg/L |
|---|---|---|---|---|---|---|
| 1 | IOS | Bright | Cuvette | DI | 461 | 0.01–1.0 |
| 2 | | | Bottle | DI | 168 | 0.01–1.0 |
| 3 | | | | Spiked | 48 | 0.2–0.9 |
| 4 | | | Dish | DI | 174 | 0.01–1.0 |
| 5 | | | | Spiked | 48 | 0.2–0.9 |
| 6 | | | Cuvette | Real 1 | 96 | 0–0.12 |
| 7 | | | | Real 2 | 102 | 0–0.13 |
| 8 | | Dim | | DI | 105 | 0.01–1.0 |
| 9 | | Dark | | DI | 108 | 0.01–1.0 |
| 10 | Android | Bright | Cuvette | DI | 174 | 0.01–1.0 |
| 11 | | | Bottle | DI | 168 | 0.01–1.0 |
| 12 | | | | Spiked | 48 | 0.2–0.9 |
| 13 | | | Dish | DI | 174 | 0.01–1.0 |
| 14 | | | | Spiked | 48 | 0.2–0.9 |

Note: the water type is the sample solution, where "DI" means nano-pure water; "Spiked" means Lake Erie water spiked to different P concentrations; "Real" 1 or 2 means the first or second time of in-situ sampled water.

image; and then R, G, and B values were recorded separately as three input features. The two phone systems were set as 0 (IOS) and 1 (Android); three light conditions were set as 0 (bright), 1 (dark), and 2 (dim); and three containers were set as 0 (cuvette), 1 (bottle), and 2 (dish) for model input. All those features are recorded in the supporting data along with the corresponding concentrations and absorbance, and the Python code template is available freely online: https://github.com/cwrukaizhang/phosphorus_detection.git. All aforementioned inputs were considered at the very beginning of the model development; then feature importance was discussed and excluded sequentially to build new models. Simplifying the model inputs could reduce the model complexity and increase the applicability of the P estimation method.

As the primary objective of this study was to develop an easy and effective method for P measurement, tailored for citizen science, real water samples were collected twice in the great Cleveland area from a sewage plant, Lake Erie, local creeks, ponds, and pits in farmlands to test the method, sampling sites see Fig. S4. Three surface water samples were collected using 50-mL glass bottles from publicly accessible watersides. The sampling dates were March 30th and April 30th, 2023, and the collected samples were analyzed on the same day. Both field and experimental samples were filtered through 0.22 μm pore-size filters for P measurement. For the application of the method, data from the first round of sampling were included in the training set along with all other laboratory-generated data (DI and spiked). The data from the second round of sampling were set as the external test set to evaluate the method. Due to the low concentrations in the field samples, they were further spiked to concentrations ranging from 0.2 to 0.9 mg P/L at an interval of 0.1 mg/L. The experimental setup and operations are illustrated in Fig. 3. All the sample groups are labeled and summarized in Table 3.

### *Machine learning models*

All the collected images were randomly stratified split to training and test sets in a 4:1 ratio across the 14-labeled groups for 20 times by changing the random state from 42 to 61. To select the best ML algorithm(s) for the modeling purposes, we compared the performance of five widely used ML algorithms: bootstrap aggregating (BA) (Breiman, 1996), extreme tree (ET) (Geurts et al., 2006), gradient boosting (GB) (Friedman, 2001), RF, an ensemble learning method that builds many trees with randomized features for classification and regression tasks (Breiman, 2001; Hastie et al., 2009), and XGBoost (XGB) (Chen and Guestrin, 2016), where light, phone, container, and three individual values of RGB channels were used as the model input to estimate the P concentrations, see the supporting data. In addition, grid search was employed for hyperparameter optimization with the data randomly stratified split for the algorithms mentioned above during the model training. All 5 model hyperparameters were optimized by changing the number of trees from 10 to 100 in a step of 10. Additionally, we changed the learning rate for GB from 0.05, 0.1, 0.5, to 1; the learning rate for XGB from 0.05, 0.1, 0.5, to 1, and the max depth from 5, 10, 20, 50, 80, to 100. A convolutional neural network (CNN) was also employed for comparison, since CNN was widely applied in image recognition studies, such as algal genera detection (Krause et al., 2020) and algal bloom monitoring (Pyo et al., 2020). The CNN model was directly applied to images without other model inputs and optimized by changing the number of epochs from 10 to 50 in a step of 10. For comparison, an empirical statistical model, the multilinear regression (MLR) model, was applied to build the relationship between RGB values and the known P concentrations as the baseline model.

The obtained regression models were evaluated by the $R^2$ and root mean squared error (RMSE) values. The P limit of detection was calculated based on the 3 times the standard deviation of the estimated P concentrations when the actual P concentration was 0 (Heidari-Bafroui et al., 2021).

RF classification models were further developed to evaluate the P extent in water samples. With 0.05 or 0.1 mg P/L as the threshold for level 1 and level 2 (as the suggested limits are 0.05 mg P/L in streams entering lakes and 0.1 mg P/L in running waters (Litke, 1999)), P concentrations from 0 to 1.0 mg P/L were divided into 4 levels: level 1 (L1, 0–0.05 mg P/L or 0–0.1 mg P/L), level 2 (L2, 0.05–0.4 mg P/L or 0.1–0.4 mg P/L), level 3 (L3, 0.4–0.7 mg P/L), and level 4 (L4, 0.7–1.0 mg P/L). The data distribution of each level is generalized in Table S2. The classification models were evaluated by accuracy and the Kappa coefficient, a factor to examine the agreement between predictions and observations.

## CRediT authorship contribution statement

**Haiping Ai:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kai Zhang:** Writing – review & editing, Validation, Software, Methodology, Conceptualization. **Huichun Zhang:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data Availability

Data will be made available as part of the supplementary materials. Data will be made available on request.

## Acknowledgements

This work was funded by the Ohio Water Development Authority #8671.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.wroa.2024.100217.

## References

Ai, H., Zhang, K., Sun, J., Zhang, H., 2023. Short-term Lake Erie algal bloom prediction by classification and regression models. Water Res., 119710
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Chen, T. and Guestrin, C. 2016 Xgboost: a scalable tree boosting system, pp. 785–794.
Costa, D., Aziz, U., Elliott, J., Baulch, H., Roy, B., Schneider, K., Pomeroy, J., 2020. The Nutrient App: developing a smartphone application for on-site instantaneous community-based NO3 and PO4 monitoring. Environ. Modell. Softw. 133, 104829.
Das, P., Chetry, B., Paul, S., Bhattacharya, S.S., Nath, P., 2022. Detection and quantification of phosphate in water and soil using a smartphone. Microchem. J. 172, 106949.
de Souza, V.S., Silva, S.M.S., Maia, L.C., dos Santos Monteiro, D., Souza, A.L.V., 2023. Development of a methodology for phosphorus determination in soils, water, and biofertilizer using digital image analysis-DIA. Chemom. Intell. Lab. Syst. 240, 104929.
Downing, J.A., 2010. Emerging global role of small lakes and ponds: little things mean a lot. Limnetica 29 (1), 0009–0024.
U.S. EPA, 1978. Method 365.3: Phosphorus, All Forms (Colorimetric, Ascorbic acid, Two Reagent). United States Environmental Protection Agency (USEPA), Washington, DC.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42.
Girshick, R., Donahue, J., Darrell, T., Malik, J., 2015. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern. Anal. Mach. Intell. 38 (1), 142–158.
Granica, M., Tymecki, Ł., 2019. Analytical aspects of smart (phone) fluorometric measurements. Talanta 197, 319–325.
Guo, T., Johnson, L.T., LaBarge, G.A., Penn, C.J., Stumpf, R.P., Baker, D.B., Shao, G., 2020. Less agricultural phosphorus applied in 2019 led to less dissolved phosphorus transported to Lake Erie. Environ. Sci. Technol. 55 (1), 283–291.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer Science & Business Media.
Heidari-Bafroui, H., Ribeiro, B., Charbaji, A., Anagnostopoulos, C., Faghri, M., 2021. Portable infrared lightbox for improving the detection limits of paper-based phosphate devices. Measurement 173, 108607.
Ho, J.C., Michalak, A.M., 2017. Phytoplankton blooms in Lake Erie impacted by both long-term and springtime phosphorus loading. J. Great Lakes Res. 43 (3), 221–228.
Jones, Z., Linder, F., 2015. Exploratory Data Analysis Using Random Forests, Chicago, USA.
Khan, F.A., Naushin, F., Rehman, F., Masoodi, A., Irfan, M., Hashmi, F., Ansari, A.A., 2014. Eutrophication: global scenario and local threat to dynamics of aquatic ecosystems. Eutrophication: Causes, Consequences and Control: Volume 2, pp. 17–27.
Krause, L.M., Koc, J., Rosenhahn, B., Rosenhahn, A., 2020. Fully convolutional neural network for detection and counting of diatoms on coatings after short-term field exposure. Environ. Sci. Technol. 54 (16), 10022–10030.
Lavanya, V., Nayak, A., Dasgupta, S., Urkude, S., Dey, S., Biswas, A., Li, B., Weindorf, D. C., Chakraborty, S., 2023. A smartphone-integrated imaging device for measuring nitrate and phosphate in soil and water samples. Microchem. J. 193, 109042.
Li, H., Fang, T., Tan, Q.G., Ma, J., 2022. Development of a versatile smartphone-based environmental analyzer (vSEA) and its application in on-site nutrient detection. Sci. Total Environ. 838, 156197.
Li, X., Liu, B., Hu, Z., Liu, P., Ye, K., Pan, J., Niu, X., 2020. Smartphone-assisted off- on photometric determination of phosphate ion based on target-promoted peroxidase-mimetic activity of porous CexZr1-xO2 (x≥ 0.5) nanocomposites. Environ. Res. 189, 109921.
Litke, D., 1999. Review of phosphorus control measures in the United States and their effects on water quality: US Geological Survey Water-Resources Investigations Report 99-4007. US Geological Survey Fact Sheet FS-007-98, South Carolina.
McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem. Med. 22 (3), 276–282.
Michalak, A.M., Anderson, E.J., Beletsky, D., Boland, S., Bosch, N.S., Bridgeman, T.B., Chaffin, J.D., Cho, K., Confesor, R., Daloğlu, I., 2013. Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. Proc. Natl. Acad. Sci. 110 (16), 6448–6452.
Moonrungsee, N., Pencharee, S., Jakmunee, J., 2015. Colorimetric analyzer based on mobile phone camera for determination of available phosphorus in soil. Talanta 136, 204–209.
Mrdjen, I., Fennessy, S., Schaal, A., Dennis, R., Slonczewski, J.L., Lee, S., Lee, J., 2018. Tile drainage and anthropogenic land use contribute to harmful algal blooms and microbiota shifts in inland water bodies. Environ. Sci. Technol. 52 (15), 8215–8223.
Murphy, J., Riley, J.P., 1962. A modified single solution method for the determination of phosphate in natural waters. Anal. Chim. Acta 27, 31–36.
Pyo, J., Park, L.J., Pachepsky, Y., Baek, S.S., Kim, K., Cho, K.H., 2020. Using convolutional neural network for predicting cyanobacteria concentrations in river water. Water Res. 186, 116349.
Sala, L., Mujeriego, R., 2001. Cultural eutrophication control through water reuse. Water Sci. Technol. 43 (10), 109–116.
Sarwar, M., Leichner, J., Naja, G.M., Li, C.Z., 2019. Smart-phone, paper-based fluorescent sensor for ultra-low inorganic phosphate detection in environmental samples. Microsyst. Nanoeng. 5 (1), 56.
Schubert, A., Pifer, L., Cheng, J., McElmurry, S.P., Kerkez, B., Love, N.G., 2022. An Automated toolchain for camera-enabled sensing of drinking water chlorine residual. ACS ES. T. Eng. 2 (9), 1697–1708.
Thermo Scientific, (2023) Analyzing phosphate by ion chromatography. https://www.thermofisher.com/us/en/home/industrial/chromatography/chromatography-learning-center/ion-chromatography-information/ion-chromatography-analysis-analyte/analyzing-phosphate-ion-chromatography.html.April9th.
Wu, H., Ling, Y., Ju, S., Chen, Y., Xu, M., Tang, Y., 2022. A smartphone-integrated light-up lanthanide fluorescent probe for the visual and ratiometric detection of total phosphorus in human urine and environmental water samples. Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 279, 121360.
Xing, Y., Xue, B., Lin, Y., Wu, X., Fang, F., Qi, P., Guo, J., Zhou, X., 2022. A cellphone-based colorimetric multi-channel sensor for water environmental monitoring. Front. Environ. Sci. Eng. 16 (12), 155.
Zheng, S., Li, H., Fang, T., Bo, G., Yuan, D., Ma, J., 2022. Towards citizen science. On-site detection of nitrite and ammonium using a smartphone and social media software. Sci. Total Environ. 815, 152613.