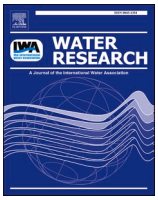
Water Research 260 (2024) 121861

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres



Using machine learning models to estimate *Escherichia coli* concentration in an irrigation pond from water quality and drone-based RGB imagery data

Seok Min Hong [a,b], Billie J. Morgan [a], Matthew D. Stocker [a], Jaclyn E. Smith [a], Moon S. Kim [a], Kyung Hwa Cho [c,**], Yakov A. Pachepsky [a,*]

[a] USDA-ARS Environmental Microbial and Food Safety Laboratory, 10300 Baltimore Ave, Bldg. 173, Beltsville, MD, 20705, USA
[b] Department of Civil Urban Earth and Environmental Engineering, Ulsan National Institute of Science and Technology, UNIST-gil 50, Ulsan, 44919, South Korea
[c] School of Civil, Environmental and Architectural Engineering, Korea University, Seoul, 02841, South Korea

ARTICLE INFO

Keywords:
Microbial water quality
Water qualoity parameters
RGB images
unmanned aerial
Vehicle
Machine learning algorithms
Escherichia coli

ABSTRACT

The rapid and efficient quantification of *Escherichia coli* concentrations is crucial for monitoring water quality. Remote sensing techniques and machine learning algorithms have been used to detect *E. coli* in water and estimate its concentrations. The application of these approaches, however, is challenged by limited sample availability and unbalanced water quality datasets. In this study, we estimated the *E. coli* concentration in an irrigation pond in Maryland, USA, during the summer season using demosaiced natural color (red, green, and blue: RGB) imagery in the visible and infrared spectral ranges, and a set of 14 water quality parameters. We did this by deploying four machine learning models – Random Forest (RF), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGB), and K-nearest Neighbor (KNN) – under three data utilization scenarios: water quality parameters only, combined water quality and small unmanned aircraft system (sUAS)-based RGB data, and RGB data only. To select the training and test datasets, we applied two data-splitting methods: ordinary and quantile data splitting. These methods provided a constant splitting ratio in each decile of the *E. coli* concentration distribution. Quantile data splitting resulted in better model performance metrics and smaller differences between the metrics for both the training and testing datasets. When trained with quantile data splitting after hyperparameter optimization, models RF, GBM, and XGB had R^2 values above 0.847 for the training dataset and above 0.689 for the test dataset. The combination of water quality and RGB imagery data resulted in a higher R^2 value (>0.896) for the test dataset. Shapley additive explanations (SHAP) of the relative importance of variables revealed that the visible blue spectrum intensity and water temperature were the most influential parameters in the RF model. Demosaiced RGB imagery served as a useful predictor of *E. coli* concentration in the studied irrigation pond

1. Introduction

Irrigation is implicated in the spread of enteric diseases. *Escherichia coli* is a crucial indicator of microbial water quality because of its prevalence and correlation with fecal contamination (Odonkor and Ampofo, 2013). Monitoring *E. coli* concentration is essential for safeguarding public health (EPA, 1989; FDA, 2023; Stocker et al., 2022). However, understanding *E. coli* dynamics in irrigation water is challenging because *E. coli* concentration varies both spatially and temporally. Therefore, there is growing interest in estimating *E. coli* concentration from readily available data. To this end, the development of accurate estimation methods based on water quality data is a feasible goal, because *E. coli* concentration is influenced by various water quality parameters (Blaustein et al., 2013; Francy et al., 2013).

Although water quality parameters have been recognized as significant predictors of *E. coli* concentration, strong correlations have rarely been observed. This is likely because simultaneous changes in numerous water quality parameters complicate the identification of simple causal links between the latter and *E. coli* concentration (Stocker et al., 2019). Nevertheless, predictions of *E. coli* concentration based on water quality parameters have been reported. Sokolova et al. (2022) developed a predictive model that used water quality parameters in Gothenburg,

* Corresponding author at: USDA-ARS Environmental Microbial and Food Safety Laboratory, 10300 Baltimore Ave, Bldg. 303, Beltsville, MD, 20705, USA.
** Corresponding author at: School of Civil, Environmental and Architectural Engineering, Korea University, Seoul, 02841, South Korea.
E-mail addresses: khcho80@korea.ac.kr (K.H. Cho), yakov.pachepsky@usda.gov (Y.A. Pachepsky).

https://doi.org/10.1016/j.watres.2024.121861
Received 11 January 2024; Received in revised form 29 May 2024; Accepted 30 May 2024
Available online 31 May 2024

Sweden: they suggested that water temperature was an important proxy of the seasonal *E. coli* concentration. Seasonal factors affect *E. coli* concentration via nutrient transportation and *E. coli*'s sensitivity to solar radiation. Tousi et al. (2021) used multivariant filter feature selection and machine learning models to quantify *E. coli* concentration from pH and turbidity. The influence of pH on *E. coli* depends on environmental conditions; water turbidity affects the penetration of ultraviolet radiation, which can inactivate *E. coli* (Abdelzaher et al., 2010). Additionally, algae have been identified as significant predictors of *E. coli* concentration. Stocker et al. (2022) predicted *E. coli* concentration from the concentration of algal pigments such as chlorophyll-a and phycocyanin: *E. coli* interact with algae by supplying organic nutrients and blocking solar radiation.

However, traditional methods for predicting *E. coli* concentration from water quality parameters involve on-site sampling. These methods are resource-intensive, and can be prohibitively expensive because of equipment, supplies, transportation, labor, and other costs (Jin et al., 2017). One practical approach bypassing these limitations is the utilization of small, unmanned aircraft system (sUAS) imagery to infer *E. coli* concentration. sUAS images at specific wavelengths, or their combination, can be used to estimate important attributes of aqueous bacterial habitats, such as turbidity and chlorophyll-a (Shin et al., 2024). Such images are easier and cheaper to obtain than the data compiled from traditional water quality monitoring. Images obtained with hyperspectral cameras (Hong et al., 2023; Kwon et al., 2020), and those utilizing wide wavelength ranges have also been used to characterize water quality. Flynn and Chapra (2014) used sUAS-based natural color (red, green, and blue band: RGB) imagery to map algal accumulation. Because RGB imagery can potentially register differences in *E. coli* habitats, it emerges as a promising tool for predicting *E. coli* concentration in water. This tool warrants further exploration. To date, remote sensing images, as well as water quality parameters that characterize site-specific microbial habitats, have been explored most effectively by using machine learning (Stocker et al., 2022; Weller et al, 2021).

Machine learning models have demonstrated outstanding performance with multiple potential predictors (Kim et al., 2023; Ma et al., 2021). The advantage of machine learning algorithms is that they can quantify predicted target outputs in the absence of predefined mathematical equations that relate the input variables to the target variable of interest (Abbas et al., 2023). This property of machine learning modeling is especially useful when the physical processes that enable the formulation of process-based models are poorly understood (Nguyen et al., 2023; Thomas et al., 2018). Machine learning models such as Random Forest (RF) and Extreme Gradient Boosting (XGB), have been utilized to predict water quality parameters in inland waters (Brooks et al., 2016; Jeong et al., 2024; Weller et al., 2021b; Zhu et al., 2022).

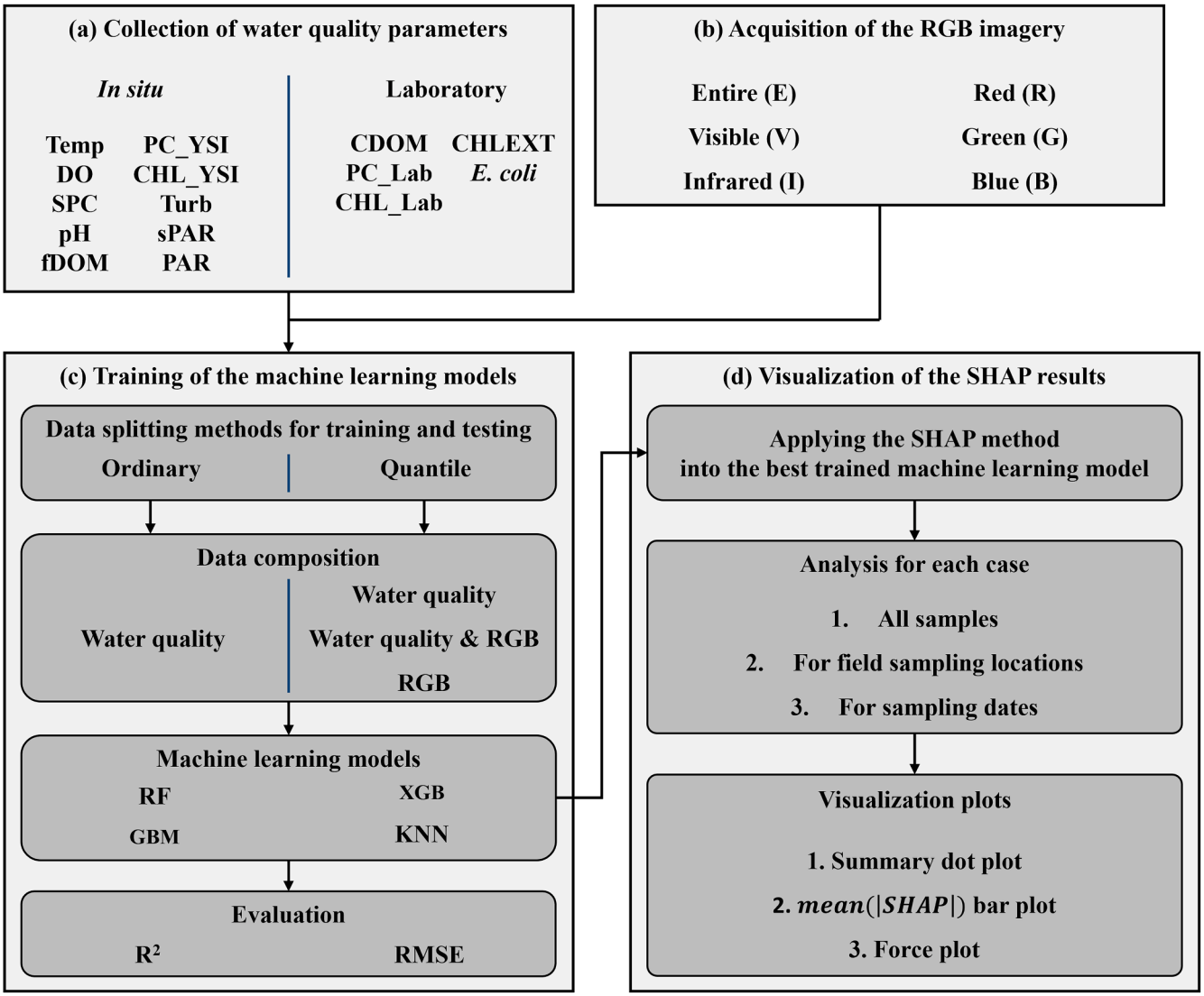


Fig. 1. Flowchart for data collection, processing, and the application and evaluation of machine learning models.

Predictive model building with machine learning algorithms is usually accompanied by the ranking of input variables according to the strength of their influence on the target variable, with a variety of metrics used to characterize this strength (Wei et al., 2015; Park et al., 2022).

This study (a) evaluates the efficacy of RGB imagery as the sole or complementary source of information for predicting *E. coli* concentration in a water body (irrigation pond), (b) compares the performance of machine learning models when water quality parameters were used alone with that when the latter were used in combination with RGB data, and (c) characterizes the strength of RGB imagery and water quality parameters as predictors of *E. coli* concentration. By doing so, this study advances the application of RGB data and machine learning models in water quality assessment. By speeding up the water quality assessment process while reducing its cost, the approach developed here can contribute to public health and landscape management.

2. Materials and methods

The research process consisted of four parts (Fig. 1): (a) collection of water quality data, (b) acquisition and processing of RGB images, (c) training and evaluation of four machine learning models using two data-splitting methods, and (d) visualization of the relative contribution of each variable to model performance.

2.1. Study site and water quality parameters

This study was conducted at the Wye Research and Education Center of the University of Maryland, USA. There, an excavated pond, approximately 3780 m^2 in area and 2.7 m-deep on average, is used to irrigate the surrounding farmland (Fig. 2). The pond is recessed in the landscape and receives runoff from cropland and farm parking lots. Runoff also enters at the northern end of the pond through an ephemeral stream that drains a forested area. The pond outflow is located at the southern end and drains only when the water level is high. Before our sampling (in summer 2018), the adjacent farmland had received inorganic fertilizers in March, but no animal manure had been applied.

We selected 34 sampling locations in a grid pattern across the pond surface (Fig. 2). Twenty-four sites were located nearshore; the remaining ten points were located in the interior of the pond. Water samples were collected at a depth of 0–15 cm over a two-hour period each time, on June 6, July 10, August 7, August 23, and September 20, 2018. Pond-interior point samples were obtained using a small boat. Nearshore sampling was performed using a 500 mL plastic dipper attached to a 1.5 m-long handle. Immediately after collection, the water samples were placed on ice and transported to the laboratory.

As part of the water sampling, environmental variables were measured *in situ* using a handheld YSI EXO-2 Sonde (Xylem Inc., Yellow Springs, Ohio, USA). These measurements included temperature (Temp,°C), dissolved oxygen (DO, mg/L), pH, specific conductance (SPC, μS/cm), fluorescent dissolved organic matter (fDOM, ppb), cyanobacteria phycocyanin (PC_YSI, relative fluorescence units: RFU), *in situ* chlorophyll-a (CHL_YSI, RFU), and turbidity (Turb, Nephelometric Turbidity Units: NTU). The photosynthetic active radiation (PAR, $W{\cdot}m^{-2}$) was measured above the water surface at the time of sampling, while the submerged PAR (sPAR, $W{\cdot}m^{-2}$) was measured under the water surface at near-0 cm depth (Apogee Instruments Inc., Logan Utah). The water quality parameters were measured from 9 to 11 am, immediately after the conclusion of the sUAS-based image collection.

The concentration of colored dissolved organic matter (CDOM), chlorophyll (CHL_Lab), and phycocyanin (PC_Lab) was determined fluorometrically in the laboratory using a benchtop analyzer (Turner Designs, CA, USA). The *E. coli* was enumerated using membrane filtration and modified thermotolerant *E. coli* (mTEC) agar (Difco, Sparks, MD, USA). The incubation time and colony counting followed the manufacturer's specifications.

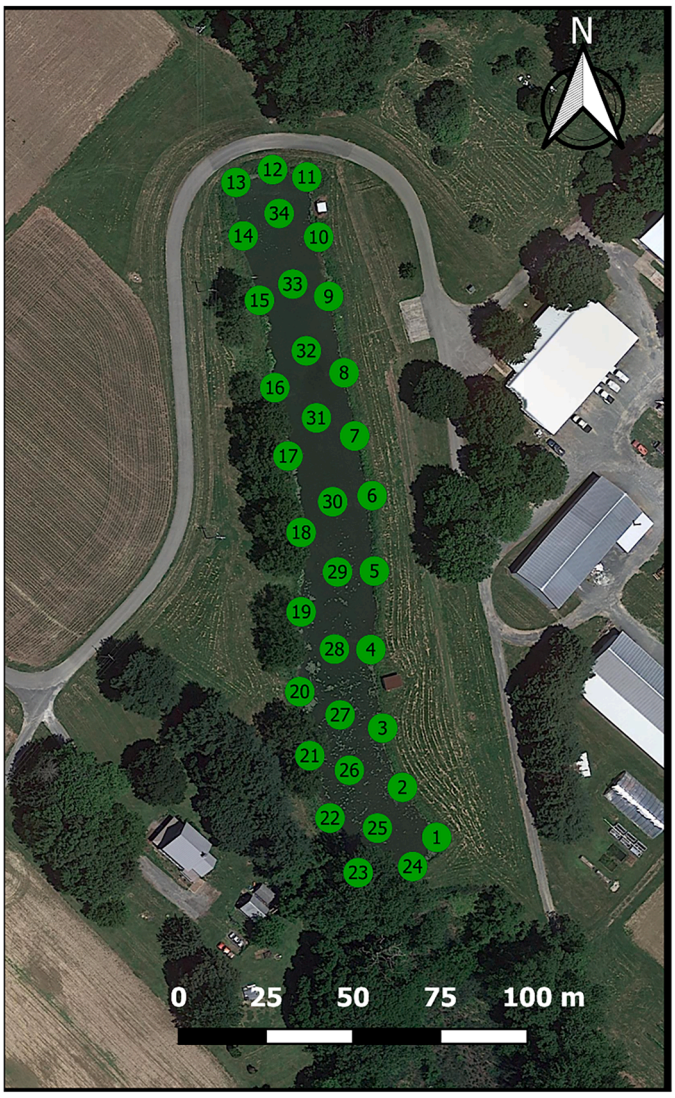


Fig. 2. Study area and sampling locations across the study pond.

2.2. Imagery acquisition and processing

The study utilized a 3DR Solo® sUAS drone (3DR, Berkeley, CA) equipped with three GoPro cameras (GoPro, San Mateo, CA, Stuntcams Inc., Ada, MI, USA). The camera lenses were modified to reduce distortion and provide a wavelength range wider than the visible spectrum (Morgan et al., 2020). The first GoPro camera captured the entire wavelength range (E), including the visible (V) and infrared (I) ranges. The second and third GoPro cameras had filters that provided V-only and I-only ranges. The images obtained by each GoPro camera were demosaiced (Kimmel, 1999) into red (entire red: ER; infrared red: IR; visible red: VR), green (entire green: EG; infrared green: IG; visible green: VG), and blue (entire blue: EB; infrared blue: IB; visible blue: VB) components. Demosaicing thus created nine data layers from the GoPro imagery data (i.e., RGB images from the three cameras) for each imaging day.

The sUAS flights were conducted at an altitude of 400 ft (approximately 130 m) under clear skies, with similar solar radiation levels. The solar zenith angle ranged from 20 to 30°. Using the georeferencing tool in ArcMap 10.5, the images were aligned with a reference image of the pond obtained from the base map of the United States Geological Survey. A 1 m diameter circular buffer was established around each water sampling location. Owing to potential changes in pond depth,

surrounding vegetation, and manual image alignment, the nearshore sampling locations may have varied between sampling days. To accommodate this variance, the 1 m buffers were incrementally shifted towards the pond's interior until they were entirely within the water-inundated area. Buffer relocation was performed using Python. Once all the buffer locations (interior and nearshore) were determined, the "clip" tool in ArcMap was used to extract the buffered portions from each of the 28 images. The "summary statistics" tool was then used to calculate the mean digital number (MDN) for each case. Several outliers within the clipped areas were removed, and the digital number distributions were found to be normal ($p > 0.05$). The MDNs were altered by less than 1 % due to the removal of the outliers. Some water sensing and sampling locations were covered by more than one image, obtained from consecutive flights. In these cases, the water quality data and RGB information from each image were treated as independent datasets.

2.3. Data splitting methods

Three of the four machine learning methods in this study required the splitting of the data into training and testing subsets in an 8:2 ratio. Ordinary (traditional) data splitting employs random sampling without considering the statistical distribution of a target variable (Krawczyk, 2016). In environmental studies, datasets are commonly imbalanced because they have skewed or multimodal distributions (Dogo et al., 2021). Imbalanced data can degrade the model performance owing to the different distributions of the training and test datasets selected by random sampling (Arief et al., 2022). Notably, the degradation of model performance due to the random splitting of imbalanced data has been reported for small datasets (Jeatrakul et al., 2010; Weller et al., 2021a). Besides ordinary splitting, in this study, we also applied quantile-based data splitting: First, the splitting ratio, that is, the ratio of the number of training and testing datasets to the number of distribution quantiles, was set to 8:2, and the number of quantiles was set to 10 (the quantiles were deciles). Subsequently, the data in each quantile of the distribution were randomly split into training and testing subsets according to the splitting ratio. Quantile splitting resulted in a similarly imbalanced statistical distributions of the *E. coli* concentration in the testing and training datasets. To compare the performance of each machine learning model with ordinary and quantile splitting, we replicated each data-splitting method ten times and compared the performance-descriptive statistics (coefficient of determination: R^2). After the model performance was evaluated, we selected specific datasets among the ten replication datasets to compare the observed *E. coli* concentration with that predicted by each model after ordinary and quantile data splitting.

2.4. Machine learning models and hyperparameter optimization

Four machine learning models were utilized to estimate the *E. coli* concentration in the pond: random forest (RF), gradient boosting machine (GBM), extreme gradient boosting (XGB), and k-nearest neighbor (KNN). Each of these models was applied on sUAS-based RGB imagery (the average for each buffer) and water quality data. The number of input variables for each of the 130 datasets was 14 and 9 for the water quality and sUAS-based RGB data, respectively. The values of the hyperparameters, that is, the data-independent model parameters, were determined for each machine learning model as described below.

The RF model is an ensemble learning method that combines multiple decision trees to obtain dependable results (Breiman, 2001). Each decision tree subdivides the input variables by determining the rules for the output variables, and the RF obtains the mean predicted results from all decision trees (Wang et al., 2021). The RF exhibits robustness against overfitting and nonlinearity because it operates without assumptions about the probability distribution of the output variables (Khanal et al., 2018). In this study, we developed an RF model using the scikit-learn library in Python 3.10. The RF model was optimized using the Bayesian optimization algorithm for the hyperparameters, including the number of trees (n_estimators), maximum depth of a tree (max_depth), minimum number of samples required to split nodes (min_samples_split), and minimum number of samples required at a leaf node (min_samples_leaf).

The GBM model is also a decision-tree-based ensemble algorithm, but differs from the RF model in that it uses various decision trees as weak learners and reduces the bias and variance of model predictions by iteratively training these decision trees through a gradient descent loss function in a process known as boosting (Friedman, 2001). Unlike the RF, which relies on randomly split subsets of input data, boosting models provide a heavier weight to instances with incorrect predictions sequentially (Krishnaraj and Honnasiddaiah, 2022). In addition, the GBM minimizes the error between observations and predictions by optimizing the gradient descent loss function during each iteration (Otchere et al., 2022). The GBM was implemented using the GradientBoostingRegressor function in the Scikit-learn library. The utilized hyperparameters were n_estimators, max_depth, min_samples_split, and learning rate (lr).

The XGB model is a unique implementation of the GBM based on regression trees (Chen and Guestrin, 2016). The XGB builds decision trees in parallel instead of sequentially. Compared to the GBM model, the XGB reduces overfitting and underfitting by using effective computing costs (Mokhtar et al., 2022). To develop the XGB model in this study, we utilized the xgboost library in Python 3.10. Hyperparameter optimization was performed for n_estimators, max_depth, lr, the maximum number of nodes (max_leaves), and the minimum sum of instance weights needed in a child (min_child_weight).

The KNN model is a supervised machine learning model that determines the k-nearest data points to a new input data point in the training dataset (Zhang et al., 2018). The model predicts new data points by calculating the average of the output values for the k-nearest neighbors. For large datasets, the KNN requires numerous calculations to acquire the distances between the training and new data points; however, this model is flexible for various data distributions (Juna et al., 2022). The KNN model was constructed using the scikit-learn library. Its hyperparameters were the number of neighbors (n_neighbors), leaf size (leaf_size), and the power parameter for the metric (p).

For each machine learning model, the hyperparameters were optimized using a Bayesian optimization algorithm built into the Bayes _opt library in Python 3.10. In addition, k-fold cross-validation, supplied by the KFold function in the Scikit-learn library, was applied to minimize model overfitting and improves the generalization performance by controlling the number of subsets (Sultana et al., 2022). The dataset was divided into five subsets, and each model was validated using one selected subset. After using the other subsets for model training, this process was repeated. The hyperparameters for each machine learning model are summarized in Supplementary Table S1.

2.5. Model performance evaluation

The four machine learning models were trained to determine the minimum loss value. The utilized loss function was the mean squared error (MSE), calculated from the observed and predicted *E. coli* concentration, as follows:

$$MSE = \frac{\sum_{i=1}^{n} \left(y_{o,i} - y_{p,i} \right)^2}{n}, \quad (1)$$

where n is the number of samples, and y_o and y_p are the observed and predicted *E. coli* concentration, respectively (CFU/100mL). After training, the models were evaluated using the coefficient of determination (R^2) and root mean square error (RMSE). These were calculated as follows:

$$R^2 = \frac{\left(\sum_{i=1}^{n}\left(y_{o,i} - \overline{y_o}\right)\left(y_{p,i} - \overline{y_p}\right)\right)^2}{\sum_{i=1}^{n}\left(y_{o,i} - \overline{y_o}\right)^2 \sum_{i=1}^{n}\left(y_{p,i} - \overline{y_p}\right)^2}, \quad (2)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{o,i} - y_{p,i}\right)^2}{n}}, \quad (3)$$

where $\overline{y_o}$ and $\overline{y_p}$ are the mean of the observed and predicted *E. coli* concentration, respectively.

2.6. Assessing the relative importance of variables: Shapley Additive Explanations (SHAP)

The SHAP method quantifies the relative importance of input variables in machine learning models (Arrieta et al., 2020; Lundberg and Lee, 2017). In this study, we applied the SHAP method to interpret the importance of the input variables used in the four machine learning models. The SHAP method applies game theory to calculate the relative contribution of each input variable (Seyrfar et al., 2021): this relative contribution is expressed by a metric known as the SHAP value. The SHAP value explains the output results of a machine learning model by considering an entire set of samples (global interpretation) or a specific sample (local interpretation). The global interpretation shows the contribution of each input variable and the sign (positive or negative) of its influence for all samples. In this study, we used the SHAP method to assess the overall influence of each input variable in the model predictions of *E. coli* concentration (global interpretation), for both the training and test datasets. In addition, we applied local interpretation to compare the selected samples for different dates and locations of the sampling points. The global and local interpretations were visualized as summary and force plots, respectively, that display the relative importance of each input variable and its relationship with the predicted outputs. To produce these plots, we used the summary and force plot functions in the SHAP library in Python 3.10.

Table 1
Descriptive statistics of the measured water quality parameters and RGB imaging.

	Variable	Unit	Min	Max	Mean
WQ	Temp	°C	25.0	35.3	28.4
	DO	mg/L	5.8	21.8	12.6
	SPC	μS/cm	10.1	164.1	143.2
	pH	-	6.6	9.2	7.9
	Turb	NTU	1.8	47.0	13.2
	PC_YSI	RFU	0.6	34.3	4.2
	CHL_YSI	RFU	1.4	166.1	13.9
	fDOM	ppb	13.8	50.6	35.2
	CHLEXT	μg/L	8.3	523.8	155.0
	CDOM	μg/L	93.5	642.8	169.2
	CHL_Lab	RFU	155.7	8715.0	1007.3
	PC_Lab	μg/L	10.9	460.2	115.3
	sPAR	$W{\cdot}m^{-2}$	55.0	1825.0	1106.1
	PAR	$W{\cdot}m^{-2}$	48.0	2000.0	1399.7
	E. coli	CFU/100mL	0.0	53.0	7.7
Remote Sensing	EB	-	22.5	103.5	59.7
	EG	-	17.7	102.9	54.8
	ER	-	47.0	191.0	109.9
	VB	-	31.3	134.7	72.9
	VG	-	26.3	149.5	84.4
	VR	-	27.7	141.3	79.7
	IB	-	52.3	149.3	83.9
	IG	-	25.6	75.7	37.3
	IR	-	84.8	191.6	108.0

3. Results

3.1. Overview of the dataset

The acquired data are presented in Table 1. The water temperature increased from June 26th to August 7th, and then decreased. The highest average water temperature (>30 °C) occurred on August 7th. In the other dates, the water temperature ranged from 25 to 30 °C. Turb, PC_YSI, CHL_YSI, CHLEXT, CDOM, CHL_Lab, and PAR attained their highest values on August 23rd and their lowest values on September 20th. The highest *E. coli* concentration (53.0 CFU/100mL) was measured on September 20th at the bank side. The mean and standard deviation of *E. coli* concentration was 4.5 ± 2.2, 5.3 ± 5.5, 1.8 ± 2.6, 10.3 ± 7.0, 43.0 ± 8.6 CFU/100mL, for sampling days 1–5, respectively. In July and August, the *E. coli* concentration was 1.5 times higher than the interquartile range of the entire dataset. The Temp, DO, and PAR parameters showed relatively large variation on the same date during July–August. In addition, most water quality parameters in July and August exhibited extreme values. Among the RGB data, the ER and IR images presented higher values than the other images, while the IG images had the lowest mean value (37.337).

To assess the imbalance in the *E. coli* concentration data, we visualized the entire *E. coli* dataset (Fig. 3). In most samples, the *E coli* concentration was under 34.5 CFU/100mL; it exceeded 20.5 CFU/100mL in only eight samples. Three samples had a relatively high *E. coli* concentration (40.0 CFU/100 mL). The R^2 values for the ten replicated datasets split with ordinary and quantile data splitting are summarized in Table 2. The ordinary and quantile data splitting performed similarly well for the training datasets. The RF, GBM, and XGB models exhibited mean R^2 values greater than 0.80. However, the mean R^2 values for the test datasets were lower than 0.34. Although the maximum R^2 values for the test datasets were higher than 0.80 (except for the KNN model), the standard deviation of R^2 for ordinary data splitting was substantially higher than that for quantile data splitting. In contrast, quantile data splitting had a small standard deviation for R^2 values below 0.18. To assess the model performance with ordinary and quantile data splitting, we selected ordinary and quantile datasets with skewed and uniform training data. With ordinary data splitting, the minimum *E. coli* concentration in the training and test datasets was the same (0 CFU/100mL), while the maximum *E. coli* concentration was 19.5 and 53.0 CFU/100 mL, respectively. This specific training dataset was selected for CDF values within 90 % using the ordinary data-splitting method; however, the test dataset included a wide range of *E. coli* concentration.

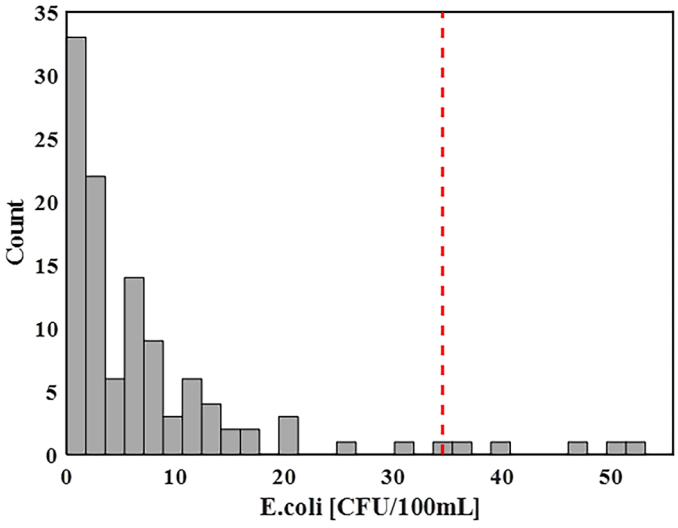


Fig. 3. *E. coli* data imbalance. The red dashed line indicates the CDF value of 95 %.

Table 2
Training-testing dataset splits by the ordinary and quantile method: R^2 values for ten replications.

R^2 for ten replication datasets		Ordinary				Quantile			
		Min	Max	Mean	Std¶	Min	Max	Mean	Std¶
RF	Training	0.731	0.865	0.815	0.043	0.745	0.897	0.835	0.046
	Test	0.055	0.876	0.344	0.265	0.305	0.736	0.590	0.142
GBM	Training	0.399	0.972	0.796	0.173	0.790	0.957	0.897	0.048
	Test	0.006	0.857	0.280	0.266	0.325	0.887	0.578	0.183
XGB	Training	0.816	0.952	0.879	0.048	0.792	0.950	0.875	0.052
	Test	0.024	0.783	0.299	0.263	0.295	0.727	0.533	0.138
KNN	Training	0.039	0.438	0.192	0.131	0.070	0.377	0.203	0.081
	Test	0.003	0.346	0.121	0.103	0.038	0.429	0.225	0.113

[*E. coli*] for selected dataset		Ordinary			Quantile		
		Min	Max	Mean	Min	Max	Mean
	Training	0.0	19.5	5.1	0.0	53.0	7.4
	Test	0.0	53.0	15.0	0.0	50.5	7.6

¶ Standard deviation

Conversely, the quantile data-splitting method retained the original distribution of the data in the training and test datasets. The minimum *E. coli* concentration was the same for the training and test datasets (0 CFU/100 mL), while the maximum *E. coli* concentration was similar (53.0 and 50.5 CFU/100 mL, respectively). The mean *E. coli* concentration was also similar for both datasets (7.4 and 7.6, respectively, for the training and test dataset).

3.2. Machine learning model performance

The performance of the four models in predicting *E. coli* concentration is summarized in Table 3. With ordinary data splitting of the training dataset consisting of water quality data only, the XGB model had the highest R^2 (0.816) and the lowest RMSE (2.0 CFU/100 mL). The RF and GBM models also performed well, with R^2 values of 0.766 and 0.776, respectively. The KNN model exhibited the poorest performance ($R^2 = 0.438$). The overall model performance for the test dataset was relatively poor, with $R^2 < 0.13$ and RMSE > 17.850 CFU/100 mL.

Fig. 4a shows an example in which the maximum *E. coli* concentration in the training dataset was 19.5 CFU/100 mL while the predicted *E. coli* concentration in the test dataset was lower than 19.5 CFU/100 mL. The XGB and KNN models, which had the best and worst performance, respectively, showed similar upper limits in the test results. The quantile data splitting of the water-quality-only dataset resulted in significant differences in performance between the training and test datasets (Table 3 and Fig. 4b). The XGB model exhibited the best performance for the training dataset, with $R^2 = 0.926$ and RMSE = 2.460 CFU/100 mL. Similarly, for the test dataset, the XGB model with quantile data splitting had an R^2 of 0.727–significantly higher than that with ordinary data splitting. The upper limit of the predicted *E. coli* concentration in the test dataset exceeded 28.0 CFU/100 mL. The RF and GBM models produced higher R^2 values (>0.689). The KNN model produced the lowest R^2 value (0.070) in the training dataset.

We compared the model performance for various input variables and their combinations, using the quantile data-splitting method. The combined water quality parameters and RGB data improved the R^2 of all machine learning models (Table 3, Fig. 5a), with the RF and XGB models exhibiting the best performance for the test dataset ($R^2 > 0.931$; RMSE < 3.9 CFU/100mL). Although the KNN model still had the lowest R^2 (0.518), the latter was significantly higher than the R^2 value obtained from water quality parameters alone ($R^2 = 0.255$). When RGB data were utilized as input variables, the machine learning models could better predict the *E. coli* concentration extremes.

The performance of the RF, XGB, and GBM models for RGB data only was lower than that for the combined water quality and RGB data (Table 3, Fig. 5b). The RF model for RGB data only had a performed slightly worse ($R^2 = 0.907$; RMSE = 5.2 CFU/100 mL) than it did for combined water quality and RGB data ($R^2 = 0.933$; RMSE = 3.9 CFU/100mL). The performance of the XGB model degraded markedly, from $R^2 = 0.931$ with combined water quality and RGB data to $R^2 = 0.773$ with RGB data only. Although most models showed a lower accuracy with RGB data only, the RF and GBM models returned remarkably similar results with and without water quality data. By contrast, the performance of the KNN model improved significantly with the combination of water quality and RGB data ($R^2 = 0.768$ for the training dataset; $R^2 = 0.929$ for the test dataset).

3.3. Sensitivity and variable importance

The relative importance of the input variables for the RF model is summarized in Fig. 6. The x-axis represents the SHAP value that indicates the effect of the predictors on the trained model outputs. The color (ranging from red to blue) represents the direction of influence (positive or negative, respectively) of the corresponding input variables. Each dot corresponds to an individual sample. In the combined water quality and RGB datasets, the VB was the most important variable, with

Table 3
Performance metrics for ordinary and quantile data splitting methods of water quality parameters only, combined water quality parameters and RGB imagery, and RGB imagery only.

Data splitting	Dataset	R^2				RMSE [CFU/100mL]			
		RF	XGB	GBM	KNN	RF	XGB	GBM	KNN
Ordinary –WQ only	Train	0.766	0.816	0.776	0.438	2.4	2.0	2.5	3.5
	Test	0.055	0.024	0.123	0.065	17.9	18.1	17.9	19.4
Quantile –WQ only	Train	0.847	0.926	0.915	0.070	3.8	2.5	2.7	8.7
	Test	0.720	0.727	0.689	0.255	7.8	7.3	7.6	11.5
Quantile – With RGB	Train	0.912	0.953	0.952	0.459	2.8	2.0	2.0	6.7
	Test	0.933	0.931	0.896	0.518	3.9	3.3	4.3	9.7
Quantile –RGB only	Train	0.905	0.949	0.932	0.768	3.0	2.0	2.4	4.4
	Test	0.907	0.773	0.872	0.929	5.2	6.7	5.9	4.7

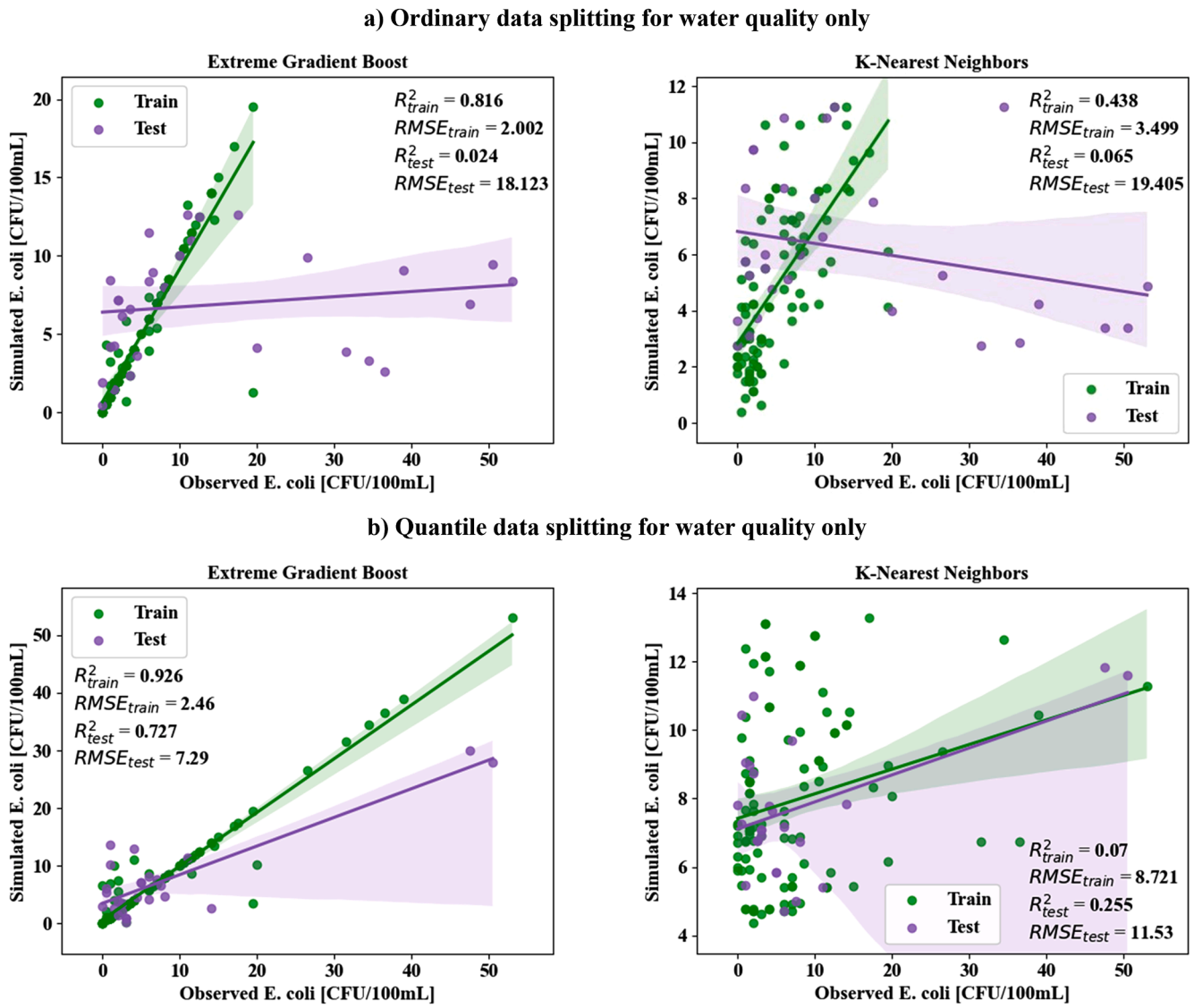


Fig. 4. Performance of the XGB and KNN models with (a) ordinary and (b) quantile data splitting for water quality data only.

a mean SHAP value of 2.128 (Fig. 6a); some samples (marked in red in Fig. 6a) had SHAP values above 17.0. Similarly, VR showed relatively high SHAP (>6.0) values. These results indicate that higher VB and VR values positively affected the predicted *E. coli* concentration. In contrast, Temp, wits a mean SHAP value of 1.985, was negatively associated with *E. coli* concentration. The pH also had a negative effect on the predicted *E. coli* concentration, but its mean SHAP value (0.812) was lower than that of Temp. When only RGB data were used, the VB was also the most important variable, with a SHAP value of 2.189 (Fig. 6b). The differences between the VG and VR variables with and without water quality data were ranked in order of importance. When only RGB data were used, the mean SHAP value for the VG and VR was 2.000 and 0.949, respectively. In the RF model, the influence of the VG data was markedly stronger than that of the other variables. The other parameters showed relatively low SHAP values (<0.800). Four parameters (VB, VG, VR, and EG) correlated positively with the predicted *E. coli* concentration.

4. Discussion

The model performance and standard deviation of the replicated datasets differed for each data-splitting method (Table 2). This may reflect the relative inefficiency of the ordinary data-splitting method in training and testing machine learning models. That is, while quantile data splitting can cover the entire data range, ordinary data splitting can only cover a high proportion of samples from a specific range of the target variable. A specific example (Fig. 4a) demonstrates the limitation of ordinary splitting owing to the artificial skewness of the training and test datasets. The range of *E. coli* concentration in the training and test datasets was 0–19.5 and 0–53.0 CFU/100 mL, respectively. In this example, however, the R^2 value for the training dataset was above 0.766–substantially higher than that for the test dataset (<0.123). With quantile splitting, by contrast, the range of *E. coli* concentration in the training and test datasets was 0–53.0 and 0–50.5 CFU/mL, respectively (Fig. 4b), and the model performance for the test dataset improved significantly (R^2 > 0.689). When the sample size is large enough and the data are adequately balanced for training a machine learning model, random data splitting can distribute the training and testing datasets proportionately. The simple random splitting of a dataset with an imbalanced distribution and a small number of samples, however, may produce a biased training dataset (An et al., 2021). The application of quantile data splitting can, therefore, enhance model performance in the case of small datasets.

With the exception of the KNN model, the machine learning models had a similar performance (Table 2, Fig. 4b). With quantile splitting, when only water quality data were used to train the machine learning models, the RF, XGB, and GBM models had R^2 values above 0.689 and

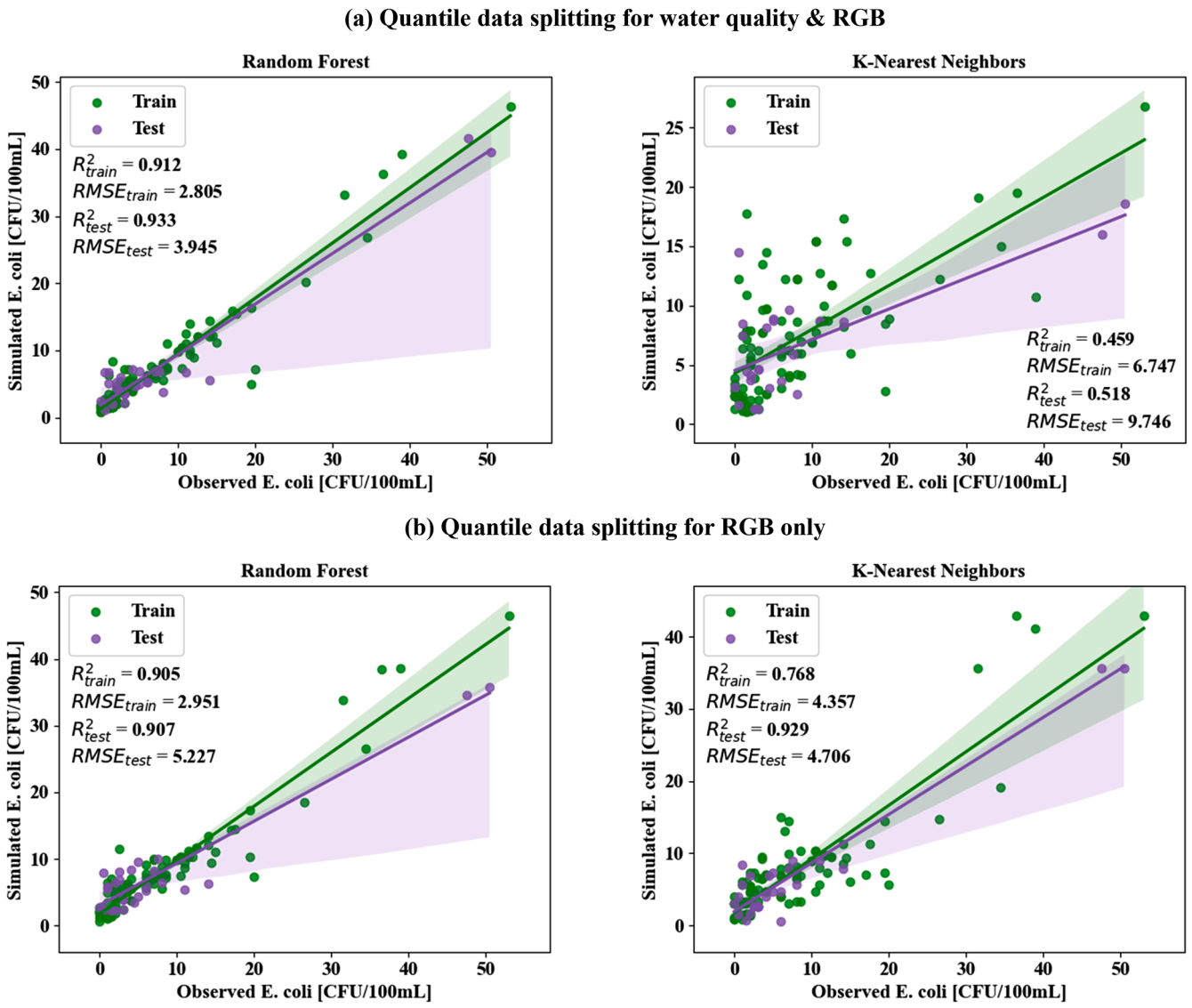


Fig. 5. Performance of the RF and KNN models with quantile data splitting for (a) water quality and RGB data, (b) RGB data only.

RMSE values below 7.8 CFU/100 mL. Among these models, the XGB exhibited the best performance, with an R^2 value of 0.926 and 0.727 for the training and test dataset, respectively. All these three models performed substantially better with the training than with the test datasets. In particular, the GBM had an R^2 value of 0.915 and 0.689 for the training and test dataset, respectively. This exemplifies the overfitting problem in machine learning exercises (Hawkins, 2004). Even though quantile data splitting was more effective than ordinary data splitting in reducing the difference in model performance between training and testing, it still produced a markedly higher R^2 for the training than for the test dataset. Similarly, the RMSE values for the training and test datasets were around 3.0 and 7.5 CFU/100 mL, respectively. Ying (2019) reported that noise can cause an overfitting problem when the training dataset is small, or its data are less representative.

The model performance was considerably better with combined water quality and RGB data than with water quality data alone (Table 2, Fig. 5a). In the training dataset, the RF, XGB, and GBM models had R^2 values above 0.912 and RMSE values below 2.8 CFU/100 mL. In the test dataset, the R^2 and RMSE values were above 0.896 and below 4.3 CFU/100 mL, respectively. Among the four models, the RF exhibited outstanding performance, with an R^2 of 0.912 for training and 0.933 for testing. Notably, the model performance for the training dataset was practically equal to that for the test dataset: the R^2 value difference between training and testing was less than 0.06. In the model runs with combined water quality and RGB data, the water quality dataset was the same as that used for the model runs with water quality data only. Therefore, the overfitting problem was due to unsuitable input variables when water-quality-only data were used. The addition of the RGB variables solved the overfitting problem and improved the prediction of *E. coli* concentration produced by the machine learning models. The improvement of machine learning model performance through the utilization of sUAS-based RGB data has been demonstrated in other studies as well (Kim and Swanson, 2018; Meyer et al., 2019). In addition, when applied to RGB-only data, the RF model yielded a similar R^2 value for the training and test datasets (an R^2 difference of just 0.002 between training and testing). This finding demonstrates the possibility that machine learning models can satisfactorily predict *E. coli* concentration from sUAS-based RGB data alone.

The KNN model exhibited the lowest R^2 values in all the three dataset scenarios (<0.52 in most scenarios: Table 2). This poor performance may have been caused by an insufficient dataset size and unsuitable input variables, similar to the overfitting problem. Ali et al. (2020) reported that the performance of a basic KNN model can be affected by data composition or size. This implies that the use of water quality

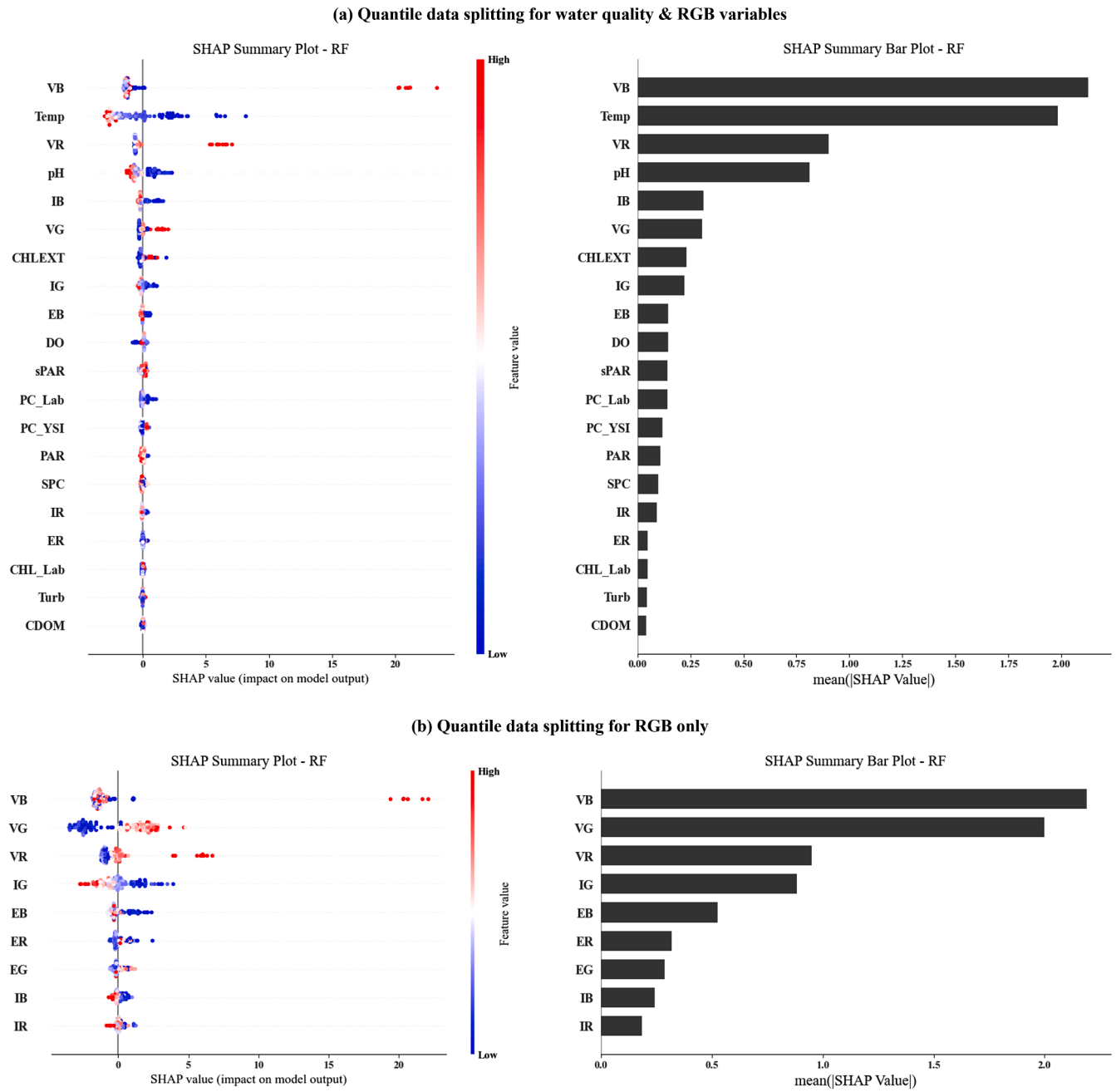


Fig. 6. SHAP summary plots for the RF model with quantile data splitting for data of (a) water quality and RGB and (b) RGB only.

parameters alone may have degraded the performance of the KNN model. Conversely, when only RGB data were applied, the R^2 value for the training and testing datasets were 0.768 and 0.929, respectively (Fig. 5b). This suggests that changes in the input parameters can improve the model's predictive performance. It is important to note that the addition of appropriate input variables is crucial for developing a better machine learning model. Adding an excessive number of variables, however, may lead to the 'curse of dimensionality' (Bowden et al., 2005; Muttil and Chau, 2007).

As shown in Fig. 6a, both temperature and pH were crucial for predicting *E. coli* concentration in the RF model. Draper et al. (2016) reported a significant correlation between *E. coli* concentration and pH, but no significant correlation between *E. coli* concentration and temperature. The influence of pH on the *E. coli* concentration in irrigation ponds has been attributed to solar inactivation caused by photo-oxidative damage to fecal microorganisms (Davies-Colley et al., 1999). Conversely, North et al. (2014) reported that temperature is a significant predictor of total coliforms, including *E. coli*. The different conclusions of these studies about the role of temperature on coliform concentration may be due to the specificities of the different study areas and seasonal factors. Stocker et al. (2021) reported a significant negative correlation between pH, temperature, and *E. coli* concentration in pond water, and attributed this correlation to the spatial and temporal variability of the measurements.

In this study, machine learning simulated a complex system of processes. *E. coli* engages in mutualistic and competitive interactions with algae, other bacteria, zooplankton, and submerged aquatic vegetation. Algae control carbon supply and nutrient availability in bacterial habitats and help *E. coli* survive by providing substrata for biofilm attachment, changing pH and the forms of oxygen, and blocking the penetration of radiation (Cho et al., 2022). Although many algal taxa release substances that inhibit bacteria (Ansa et al., 2011; Cole, 1982),

microcystin, among the most common cyanotoxins, may prolong the growth cycle of *E. coli* (Yang et al., 2008). At the same time, aqueous methanol and other organic solvents found in cyanobacteria extracts possess antibacterial properties (Falch et al., 1995; FRANKMÖLLE et al., 1992; Kellam and Walker, 1989). The water quality parameters considered in this study represent the integral synergistic effects of interactions in *E. coli* habitats. Future research should identify which set of water quality parameters and/or remote sensing data will be sufficient for obtaining a widely applicable predictive model. This model will have to be developed and tested with data from a large number of waterbodies.

To evaluate the influence of RGB data, we visualized the relationship between RGB data and model predictions of *E. coli* concentration when only RGB data were used in the models (Fig. 6b). The interpretation of this relationship is far from straightforward. However, a large body of literature has demonstrated that major *E. coli* habitat parameters, such as chlorophyll-a content, turbidity, and dissolved organic matter content, can be successfully predicted from remote and proximal sensing data. We hypothesize that, as long as these habitat parameters can be reasonably estimated, *E. coli* concentration could be predicted effectively from remote sensing data. Phytoplankton populations and light penetration are additional factors influencing *E. coli* survival and should be parameterized in *E. coli* habitats. Algae can harm bacteria since increased DO levels due to photosynthesis can lead to photooxidation; at the same time, they can favor bacterial growth by attenuating light penetration. In natural ponds, the penetration of solar radiation decreases with increasing algal biomass (Van der Steen et al., 2000). UV-B radiation (290–320 nm) causes direct DNA damage in *E, coli* (Shilton, 2006); typically, however, this damage only occurs in the top few centimeters of the water column because of the presence of surface-water algal biomass attenuates solar radiation with depth. A high algal biomass can slow the inactivation of *E. coli* caused by the presence of photosensitizers in *E. coli* cells (Bolton et al., 2010). In addition, algae produce secondary metabolites that provide photoprotection (Carreto and Carignan, 2011). The inclusion of backscattering property data, as suggested by one of our reviewers, may thus be crucial for estimating bacterial populations in water from remote sensing data.

Moreover, since this study relied on data collected in a specific irrigation pond during the irrigation season (June–September), the applicability of our results is limited to the summer irrigation season. To evaluate and refine the method for estimating *E. coli* concentration developed in this work, future research should utilize data from many more waterbodies, collected over longer seasons.

5. Conclusions

This study aimed to predict *E. coli* concentrations using water quality parameters and sUAS-based RGB data, and to evaluate four machine learning models using two different data-splitting methods. We compared the performance of the models with different sets of input parameters, including water quality and RGB data, and analyzed the relative influence of each variable on model predictions using the SHAP method. The major findings of this study are as follows.

- For a small number of samples, the ordinary data-splitting method had a larger deviation in R^2 values than the quantile data-splitting method. For all machine learning models, when the training dataset had a skewed distribution, ordinary data splitting resulted in a poorer model performance with the test dataset. While most machine learning models had a relatively high R^2 value (>0.766) for the training dataset, they had a much lower R^2 value for the test dataset (<0.123). Unlike ordinary data splitting, quantile data splitting, which selects training and test samples while approximating the original distribution, provided a substantially better model performance with a test dataset that used water quality parameters only. For all machine learning models except the KNN, the R^2 value was higher (>0.689) with quantile data splitting than with ordinary data splitting.
- When the input dataset included sUAS-based RGB data, the performance of the machine learning models improved ($R^2 > 0.896$). The RF model exhibited the highest R^2 value (0.933) for the test dataset. The inclusion of RGB data improved the performance of the KNN model ($R^2 = 0.518$).
- According to the SHAP analysis, visible blue (VB) was the most important parameter in the RF model. Moreover, temperature and visible red (VR) had a relatively strong influence on the predicted *E. coli* concentration.

This study demonstrated that *E. coli* concentration can be accurately predicted using combined water quality and sUAS-based RGB data: the *E. coli* concentration for an irrigation pond in Maryland, USA, was mapped successfully using machine learning models with RGB data from the divided buffer regions. Further research in this direction is needed in order to develop a system for the efficient and cost-effective prediction of *E. coli* concentration in waterbodies.

CRediT authorship contribution statement

Seok Min Hong: Writing – original draft, Validation, Investigation, Formal analysis. **Billie J. Morgan:** Investigation, Data curation. **Matthew D. Stocker:** Methodology, Investigation, Data curation. **Jaclyn E. Smith:** Investigation, Data curation. **Moon S. Kim:** Resources, Methodology. **Kyung Hwa Cho:** Writing – review & editing, Supervision, Funding acquisition. **Yakov A. Pachepsky:** Writing – review & editing, Project administration, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare no conflict of interests.

Data availability

Data will be made available on request.

Acknowledgements

This research was partially funded by the USDA-ARS project 440973 “Improving Pre-harvest Produce Safety through Reduction of Pathogen Levels in Agricultural Environments t Development and Validation of Farm-Scale Microbial Quality Model for Irrigation Water Sources.”

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2024.121861.

References

Abbas, A., Park, M., Baek, S.-S., Cho, K.H., 2023. Deep learning-based algorithms for long-term prediction of chlorophyll-a in catchment streams. J. Hydrol. (Amst) 626, 130240.

Abdelzaher, A.M., Wright, M.E., Ortega, C., Solo-Gabriele, H.M., Miller, G., Elmir, S., Newman, X., Shih, P., Bonilla, J.A., Bonilla, T.D., 2010. Presence of pathogens and indicator microbes at a non-point source subtropical recreational marine beach. Appl. Environ. Microbiol. 76 (3), 724–732.

Ali, A., Hamraz, M., Kumam, P., Khan, D.M., Khalil, U., Sulaiman, M., Khan, Z., 2020. A k-nearest neighbours based ensemble via optimal model selection for regression. IEEe Access. 8, 132095–132105.

An, C., Park, Y.W., Ahn, S.S., Han, K., Kim, H., Lee, S.-K., 2021. Radiomics machine learning study with a small sample size: single random training-test set split may lead to unreliable results. PLoS. One 16 (8), e0256152.

Ansa, E., Lubberding, H.J., Ampofo, J., Gijzen, H., 2011. The role of algae in the removal of *Escherichia coli* in a tropical eutrophic lake. Ecol. Eng. 37 (2), 317–324.

Arief, H.A.a., Thomas, P.J., Wiktorski, T., 2022. Better modeling out-of-distribution regression on distributed acoustic sensor data using anchored hidden state mixup. IEEe Trans. Industr. Inform. 19 (1), 296–305.
Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., 2020. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.
Blaustein, R., Pachepsky, Y., Hill, R., Shelton, D., Whelan, G., 2013. *Escherichia coli* survival in waters: temperature dependence. Water. Res. 47 (2), 569–578.
Bolton, N.F., Cromar, N.J., Hallsworth, P., Fallowfield, H.J., 2010. A review of the factors affecting sunlight inactivation of micro-organisms in waste stabilisation ponds: preliminary results for enterococci. Water Sci. Technol. 61 (4), 885–890.
Bowden, G.J., Dandy, G.C., Maier, H.R., 2005. Input determination for neural network models in water resources applications. Part 1—background and methodology. J. Hydrol. (Amst) 301 (1-4), 75–92.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Brooks, W., Corsi, S., Fienen, M., Carvin, R., 2016. Predicting recreational water quality advisories: a comparison of statistical methods. Environ. Model. Softw. 76, 81–94.
Carreto, J.I., Carignan, M.O., 2011. Mycosporine-like amino acids: relevant secondary metabolites. Chemical and ecological aspects. Mar. Drugs 9 (3), 387–446.
Chen, T. and Guestrin, C. 2016 Xgboost: a scalable tree boosting system, pp. 785-794.
Cho, K.H., Wolny, J., Kase, J.A., Unno, T., Pachepsky, Y., 2022. Interactions of *E. coli* with algae and aquatic vegetation in natural waters. Water. Res. 209, 117952.
Cole, J.J., 1982. Interactions between bacteria and algae in aquatic ecosystems. annual review of ecology. Evol. Syst. 13 13, 291–314. Volume.
Davies-Colley, R., Donnison, A., Speed, D., Ross, C., Nagels, J., 1999. Inactivation of faecal indicator micro-organisms in waste stabilisation ponds: interactions of environmental factors with sunlight. Water. Res. 33 (5), 1220–1230.
Dogo, E.M., Nwulu, N.I., Twala, B., Aigbavboa, C., 2021. Accessing imbalance learning using dynamic selection approach in water quality anomaly detection. Symmetry. (Basel) 13 (5), 818.
Draper, A.D., Doores, S., Gourama, H., LaBorde, L.F., 2016. Microbial survey of Pennsylvania surface water used for irrigating produce crops. J. Food Prot. 79 (6), 902–912.
Epa, U., 1989. Drinking water: national pri-mary drinking water regulations; total coliforms (Including Fecal Coliforms and *E. coli*): final rule. Fed. Regist. 54, 27544–27568.
Falch, B.S., König, G.M., Wright, A.D., Sticher, O., Angerhofer, C.K., Pezzuto, J.M., Bachmann, H., 1995. Biological activities of cyanobacteria: evaluation of extracts and pure compounds. Planta Med. 61 (04), 321–328.
FDA 2023 The New FDA Food Modernization Act (FSMA): produce safety rules. (FDA), U.F.A.D.A. (ed).
Flynn, K.F., Chapra, S.C., 2014. Remote sensing of submerged aquatic vegetation in a shallow non-turbid river using an unmanned aerial vehicle. Remote Sens. (Basel) 6 (12), 12815–12836.
Francy, D.S., Stelzer, E.A., Duris, J.W., Brady, A.M., Harrison, J.H., Johnson, H.E., Ware, M.W., 2013. Predictive models for *Escherichia coli* concentrations at inland lake beaches and relationship of model variables to pathogen detection. Appl. Environ. Microbiol. 79 (5), 1676–1688.
Frankmölle, W.P., Larsen, L.K., Caplan, F.R., Patterson, G.M., Knübel, G., Levine, I.A., Moore, R.E, 1992. Antifungal cyclic peptides from the terrestrial blue-green alga Anabaena laxa I. Isolation and biological properties. J. Antibiot. (Tokyo) 45 (9), 1451–1457.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Hawkins, D.M., 2004. The problem of overfitting. J. Chem. Inf. Comput. Sci. 44 (1), 1–12.
Hong, S.M., Abbas, A., Kim, S., Kwon, D.H., Yoon, N., Yun, D., Lee, S., Pachepsky, Y., Pyo, J., Cho, K.H., 2023. Autonomous calibration of EFDC for predicting chlorophyll-a using reinforcement learning and a real-time monitoring system. Environ. Model. Softw. 168, 105805.
Jeatrakul, P., Wong, K.W., Fung, C.C., 2010. Classification of Imbalanced Data By Combining the Complementary Neural Network and SMOTE Algorithm. Springer, pp. 152–159 pp.
Jeong, H., Park, S., Choi, B., Yu, C.S., Hong, J.Y., Jeong, T.-Y., Cho, K.H., 2024. Machine learning-based water quality prediction using octennial in-situ Daphnia magna biological early warning system data. J. Hazard. Mater. 465, 133196.
Jin, Q., Lyu, H., Shi, L., Miao, S., Wu, Z., Li, Y., Wang, Q., 2017. Developing a two-step method for retrieving cyanobacteria abundance from inland eutrophic lakes using MERIS data. Ecol. Indic. 81, 543–554.
Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A.A., Mohamed, A., Ashraf, I., 2022. Water quality prediction using KNN imputer and multilayer perceptron. Water. (Basel) 14 (17), 2592.
Kellam, S.J., Walker, J.M., 1989. Antibacterial activity from marine microalgae in laboratory culture. Br. Phycol. J. 24 (2), 191–194.
Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., Shearer, S., 2018. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. Comput. Electron. Agric. 153, 213–225.
Kim, H.G., Cho, K.H., Recknagel, F., 2023. Time-series modelling of harmful cyanobacteria blooms by convolutional neural networks and wavelet generated time-frequency images of environmental driving variables. Water. Res. 246, 120662.
Kim, H.H., Swanson, N.R., 2018. Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. Int. J. Forecast. 34 (2), 339–354.
Kimmel, R., 1999. Demosaicing: image reconstruction from color CCD samples. IEEE Trans. Image Process. 8 (9), 1221–1228.
Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Progr. Artif. Intell. 5 (4), 221–232.
Krishnaraj, A., Honnasiddaiah, R., 2022. Remote sensing and machine learning based framework for the assessment of spatio-temporal water quality in the Middle Ganga Basin. Environ. Sci. Pollut. Res. 29 (43), 64939–64958.
Kwon, Y.S., Pyo, J., Kwon, Y.-H., Duan, H., Cho, K.H., Park, Y., 2020. Drone-based hyperspectral remote sensing of cyanobacteria using vertical cumulative pigment concentration in a deep reservoir. Remote Sens. Environ. 236, 111517.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, p. 30.
Ma, Y., Song, K., Wen, Z., Liu, G., Shang, Y., Lyu, L., Du, J., Yang, Q., Li, S., Tao, H., 2021. Remote sensing of turbidity for lakes in northeast China using Sentinel-2 images with machine learning algorithms. IEEe J. Sel. Top. Appl. Earth. Obs. Remote Sens. 14, 9132–9146.
Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. Ecol. Modell. 411, 108815.
Mokhtar, A., Elbeltagi, A., Gyasi-Agyei, Y., Al-Ansari, N., Abdel-Fattah, M.K., 2022. Prediction of irrigation water quality indices based on machine learning and regression models. Appl. Water. Sci. 12 (4), 76.
Morgan, B.J., Stocker, M.D., Valdes-Abellan, J., Kim, M.S., Pachepsky, Y., 2020. Drone-based imaging to assess the microbial water quality in an irrigation pond: a pilot study. Sci. Total Environ. 716, 135757.
Muttil, N., Chau, K.-W., 2007. Machine-learning paradigms for selecting ecologically significant input variables. Eng. Appl. Artif. Intell. 20 (6), 735–744.
Nguyen, X.C., Bui, V.K.H., Cho, K.H., Hur, J., 2023. Practical application of machine learning for organic matter and harmful algal blooms in freshwater systems: a review. Crit. Rev. Environ. Sci. Technol. 1–23.
North, R., Khan, N., Ahsan, M., Prestie, C., Korber, D., Lawrence, J., Hudson, J., 2014. Relationship between water quality parameters and bacterial indicators in a large prairie reservoir: Lake Diefenbaker, Saskatchewan, Canada. Can. J. Microbiol. 60 (4), 243–249.
Odonkor, S.T., Ampofo, J.K., 2013. *Escherichia coli* as an indicator of bacteriological quality of water: an overview. Microbiol. Res. (Pavia) 4 (1), e2.
Otchere, D.A., Ganat, T.O.A., Ojero, J.O., Tackie-Otoo, B.N., Taki, M.Y., 2022. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. J. Pet. Sci. Eng. 208, 109244.
Park, J, Lee, W.H., ae Kim, K.T., Park, C.Y., Lee, S., Heo, T.Y., 2022. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. Sci. Total Environ. 832, 155070.
Seyrfar, A., Ataei, H., Movahedi, A., Derrible, S., 2021. Data-driven approach for evaluating the energy efficiency in multifamily residential buildings. Pract. Periodic. Struct. Des. Construct. 26 (2), 04020074.
Shilton, A. 2006. Pond treatment technology.
Shin, J., Lee, G., Kim, T., Cho, K.H., Hong, S.M., Kwon, D.H., Pyo, J., Cha, Y., 2024. Deep learning-based efficient drone-borne sensing of cyanobacterial blooms using a clique-based feature extraction approach. Sci. Total Environ. 912, 169540.
Sokolova, E., Ivarsson, O., Lillieström, A., Speicher, N.K., Rydberg, H., Bondelind, M., 2022. Data-driven models for predicting microbial water quality in the drinking water source using *E. coli* monitoring and hydrometeorological data. Sci. Total Environ. 802, 149798.
Stocker, M., Pachepsky, Y., Hill, R., Sellner, K., Macarisin, D., Staver, K., 2019. Intraseasonal variation of *E. coli* and environmental covariates in two irrigation ponds in Maryland, USA. Sci. Total Environ. 670, 732–740.
Stocker, M.D., Pachepsky, Y.A., Hill, R.L., 2022. Prediction of *E. coli* concentrations in agricultural pond waters: application and comparison of machine learning algorithms. Front. Artif. Intell. 4, 768650.
Stocker, M.D., Pachepsky, Y.A., Smith, J., Morgan, B., Hill, R.L., Kim, M.S., 2021. Persistent patterns of *E. coli* concentrations in two irrigation ponds from 3 years of monitoring. Water, Air, Soil Pollut. 232, 1–15.
Sultana, N., Hossain, S.Z., Abusaad, M., Alanbar, N., Senan, Y., Razzak, S., 2022. Prediction of biodiesel production from microalgal oil using Bayesian optimization algorithm-based machine learning approaches. Fuel 309, 122184.
Thomas, M.K., Fontana, S., Reyes, M., Kehoe, M., Pomati, F., 2018. The predictability of a lake phytoplankton community, over time-scales of hours to years. Ecol. Lett. 21 (5), 619–628.
Tousi, E.G., Duan, J.G., Gundy, P.M., Bright, K.R., Gerba, C.P., 2021. Evaluation of *E. coli* in sediment for assessing irrigation water quality using machine learning. Sci. Total Environ. 799, 149286.
Van der Steen, P., Brenner, A., Shabtai, Y., Oron, G., 2000. Improved fecal coliform decay in integrated duckweed and algal ponds. Water Sci. Technol. 42 (10-11), 363–370.
Wang, F., Wang, Y., Zhang, K., Hu, M., Weng, Q., Zhang, H., 2021. Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. Environ. Res. 202, 111660.
Wei, P., Lu, Z., Song, J., 2015. Variable importance analysis: A comprehensive review. Reliab. Eng. Syst. Saf. 142, 399–432.
Weller, D.L., Love, T.M., Wiedmann, M., 2021a. Comparison of resampling algorithms to address class imbalance when developing machine learning models to predict foodborne pathogen presence in agricultural water. Front. Environ. Sci. 9, 701288.
Weller, D.L., Love, T.M., Wiedmann, M., 2021b. Interpretability versus accuracy: a comparison of machine learning models built using different algorithms, performance measures, and features to predict *E. coli* levels in agricultural water. Front. Artif. Intell. 4, 628441.
Yang, C., Wang, W., Li, D., Liu, Y., 2008. Growth and antioxidant system of *Escherichia coli* in response to microcystin-RR. Bull. Environ. Contam. Toxicol. 81, 427–431.

Ying, X., 2019. An Overview of Overfitting and its Solutions. IOP Publishing, 022022 p.
Zhang, S., Cheng, D., Deng, Z., Zong, M., Deng, X., 2018. A novel kNN algorithm with data-driven k parameter computation. Pattern. Recognit. Lett. 109, 44–54.
Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., Ye, L., 2022. A review of the application of machine learning in water quality evaluation. Eco-Environ. Health.