Electronic Supplementary Information

A field-deployable water quality monitoring with machine learning based smartphone colorimetry

Vakkas Doğan[a], Tuğba Isık[b], Volkan Kılıç[a,*] and Nesrin Horzum[c,*]

[a]Department of Electrical and Electronics Engineering Graduate Program, Izmir Katip Celebi University, Izmir, Turkey.

[b]Department of Mineral Analysis and Technologies, General Directorate of Mineral Research and Exploration (MTA), Ankara, Turkey

[c]Department of Engineering Sciences, Izmir Katip Celebi University, Turkey.

***Corresponding author**
Email addresses: volkan.kilic@ikcu.edu.tr (Volkan Kılıç),

TABLES

Table S1. Classification accuracy results for all ions with different machine learning classifiers.

	Classification Accuracy for Ions (%)						
Machine Learning Classifiers	**Ammonium**	**Arsenic**	**Carbonate**	**Chloride**	**Iron**	**Nitrate**	**Sulfate**
Linear SVM	85.95	74.86	92.50	98.66	86.63	93.93	96.48
Random Forest	61.83	61.81	81.81	98.38	64.76	82.30	81.57
Decision Tree	44.80	43.82	60.97	61.57	47.74	63.21	54.35
Nearest Neighbors (KNN)	**98.69**	**95.63**	**98.10**	**99.95**	**97.88**	**98.45**	**99.72**
Bagging Classifier	45.48	44.20	60.37	88.61	35.76	57.86	41.39
Naive Bayes	58.77	59.03	80.93	84.58	55.73	73.85	63.47
Gaussian Process	65.04	71.42	85.28	99.91	69.86	90.36	97.73
Logistic Regression	76.43	80.10	92.13	99.91	85.69	95.44	98.70
Linear Discriminant Analysis	79.52	81.11	92.22	99.91	88.37	95.60	98.47
AdaBoost	58.97	54.76	64.54	19.03	63.19	70.67	22.08
Bernoulli Naive Bayes	72.34	75.38	89.68	85.42	78.61	92.50	80.83
Extra Tree Classifier	57.46	63.44	80.74	99.49	60.90	83.41	93.84
Gradient Boosting Classifier	62.74	66.72	76.90	95.00	61.93	75.47	70.79
RBF SVM	61.83	65.44	76.94	96.01	57.31	74.47	72.08
Nu SVM	50.28	56.65	69.49	89.71	41.30	63.84	59.21
Passive Aggressive Classifier	59.64	65.09	76.81	95.41	54.21	72.36	70.60
QDA	54.40	61.69	72.87	90.96	48.92	68.34	60.46
Weighted KNN	64.56	68.74	79.44	96.57	63.43	77.70	75.09
Ensemble Boosted Tree	72.98	68.82	89.81	99.35	74.79	89.09	95.97
Ensemble Bagged Tree	93.93	86.91	95.74	99.49	95.03	97.50	99.07
Ensemble Subspace Discriminant	57.02	63.78	80.37	99.44	64.62	78.73	97.45
Coarse Tree	94.69	92.63	97.10	93.20	96.88	94.48	97.95
Ensemble RUS Boosted Tree	62.02	61.11	82.45	97.27	65.00	82.30	81.76

Table S2. Evaluation of the KNN for Ammonium in terms of precision, recall and F1-score.

Concentrations (mg/L)	Precision	Recall	F1-score
0	0.994	0.989	0.992
10	0.969	0.986	0.978
30	0.975	0.975	0.975
60	0.978	0.967	0.972
100	0.978	0.981	0.979
200	1.000	0.997	0.999
400	1.000	1.000	1.000
Average	**0.985**	**0.985**	**0.985**

Table S3. Evaluation of the KNN for Arsenic in terms of precision, recall and F1-score.

Concentrations (mg/L)	Precision	Recall	F1-score
0	0.944	0.958	0.951
0.005	0.956	0.930	0.942
0.01	0.969	0.991	0.980
0.025	0.986	0.975	0.981
0.05	0.858	0.904	0.880
0.1	0.908	0.867	0.887
0.25	0.992	0.997	0.994
0.5	0.997	0.992	0.994
Average	**0.951**	**0.952**	**0.951**

Table S4. Evaluation of the KNN for Carbonate in terms of precision, recall and F1-score.

Concentrations (mg/L)	Precision	Recall	F1-score
0	1.000	0.997	0.999
70	0.997	1.000	0.999
140	1.000	1.000	1.000
215	1.000	0.997	0.999
285	0.939	0.934	0.936
430	0.933	0.941	0.937
Average	**0.978**	**0.978**	**0.978**

Table S5. Evaluation of the KNN for Iron in terms of precision, recall and F1-score.

Concentrations (mg/L)	Precision	Recall	F1-score
0	0.956	0.932	0.944
3	0.942	0.947	0.944
10	0.944	0.950	0.947
25	0.972	0.989	0.980
50	0.989	0.989	0.989
100	0.994	0.992	0.993
250	0.997	1.000	0.999
500	1.000	0.997	0.999
Average	**0.974**	**0.974**	**0.974**

Table S6. Evaluation of the KNN for Nitrate in terms of precision, recall and F1-score.

Concentrations (mg/L)	Precision	Recall	F1-score
0	0.975	0.936	0.955
10	0.942	0.971	0.956
30	0.975	0.986	0.980
60	0.997	0.989	0.993
100	0.989	0.997	0.993
200	0.992	0.997	0.994
400	0.997	0.992	0.994
Average	**0.981**	**0.981**	**0.981**

Table S7. Results of observers in different types of illumination relative to test strips.

	Ammonium (mg/L)	Carbonate (mg/L)	Chloride (mg/L)	Nitrate (mg/L)	Illumination types
Balçova					
Observer 1	10	285	0	10	Halogen (H)
	10	285	0	10	Fluorescent (F)
	10	285	0	10	Sunlight (S)
	10	285	0	10	Halogen-Fluorescent (HF)
	10	285	0	10	Halogen-Sunlight (HS)
	10	285	0	10	Fluorescent-Sunlight (FS)
	10	285	0	10	Halogen-Fluorescent-Sunlight (HFS)
	10	285	0	10	Daylight (D)
Observer 2	10	285	0	10	H
	0	215	0	10	F
	0	215	0	10	S
	10	215	0	10	HF
	0	215	0	10	HS
	10	215	0	10	FS
	0	215	0	10	HFS
	10	215	0	10	D
Observer 3	10	285	0	10	H
	10	285	0	10	F
	10	285	0	10	S
	10	285	0	10	HF
	10	285	0	10	HS
	10	285	0	10	FS
	10	285	0	10	HFS
	10	215	0	10	D
Buca					
Observer 1	0	215	0	0	H
	0	215	0	0	F
	0	215	0	0	S
	0	215	0	0	HF
	0	215	0	0	HS
	0	215	0	0	FS
	0	215	0	0	HFS
	0	215	0	0	D
Observer 2	0	215	0	0	H
	0	215	0	0	F

	0	215	0	0	S
	0	215	0	0	HF
	0	215	0	0	HS
	0	215	0	0	FS
	0	215	0	0	HFS
	0	215	0	0	D
Observer 3	0	215	0	0	H
	0	215	0	0	F
	0	215	0	0	S
	0	215	0	0	HF
	0	215	0	0	HS
	0	215	0	0	FS
	0	215	0	0	HFS
	0	215	0	0	D
Kemalpaşa					
Observer 1	10	285	500	30	H
	10	285	500	30	F
	10	285	500	30	S
	10	285	500	30	HF
	10	285	500	30	HS
	10	285	500	30	FS
	10	285	500	30	HFS
	10	285	500	30	D
Observer 2	10	285	0	30	H
	10	285	0	30	F
	10	285	0	30	S
	10	285	0	30	HF
	10	285	0	30	HS
	10	285	0	30	FS
	10	285	0	30	HFS
	10	285	0	30	D
Observer 3	10	285	0	30	H
	10	285	0	30	F
	10	285	0	30	S
	10	285	0	30	HF
	10	285	0	30	HS
	10	285	0	30	FS
	10	285	0	30	HFS
	10	285	0	30	D

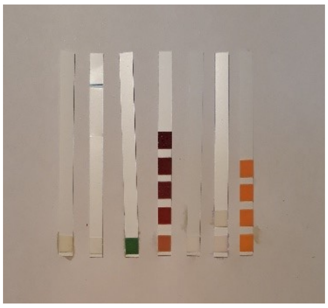
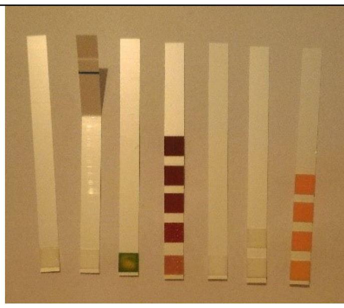
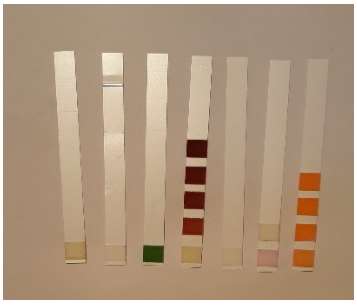
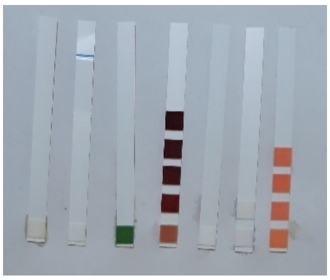
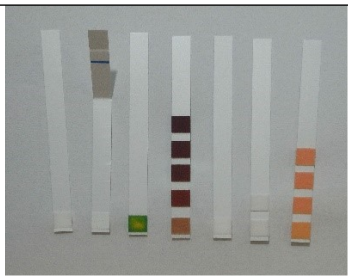
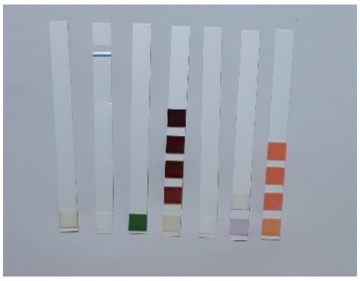
Table S8. Reference values analysed by IC

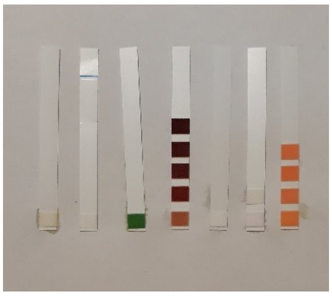
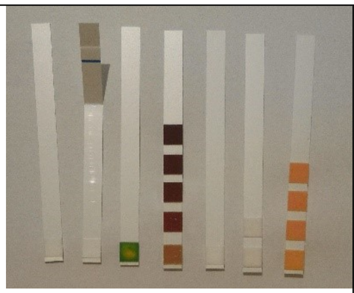
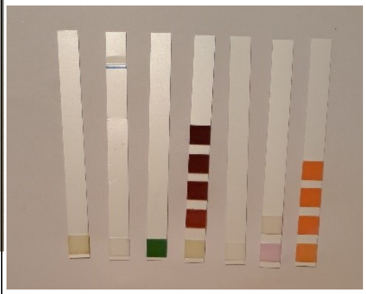
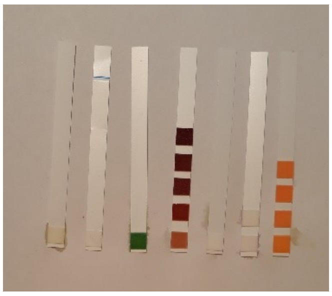
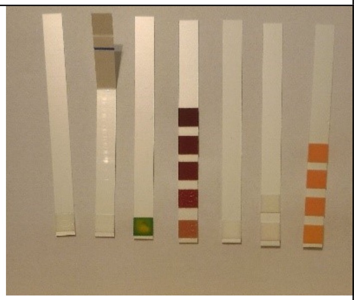
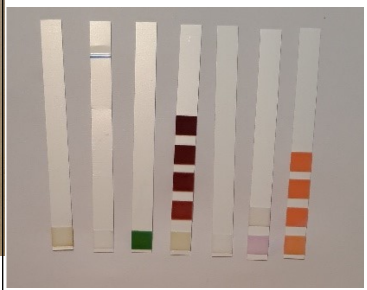
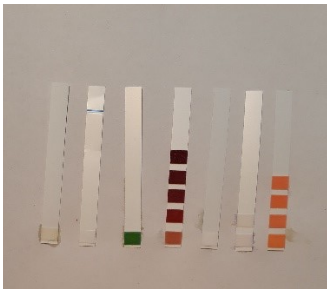
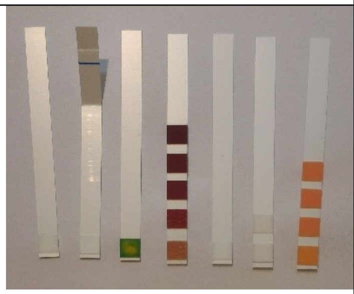
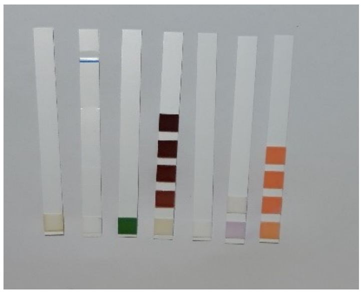
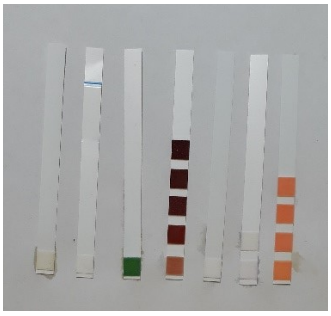
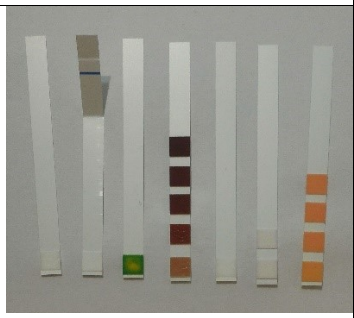
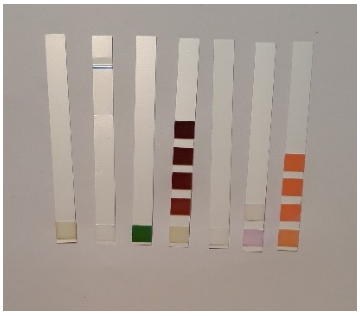
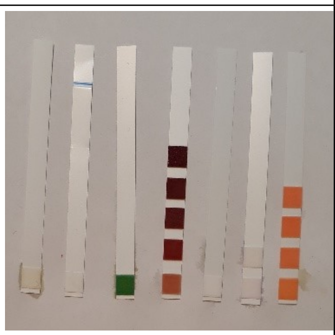
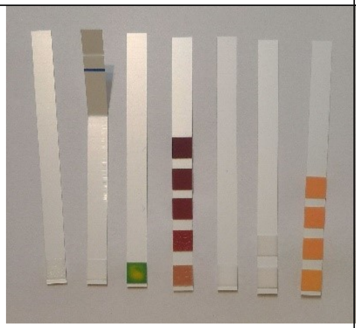
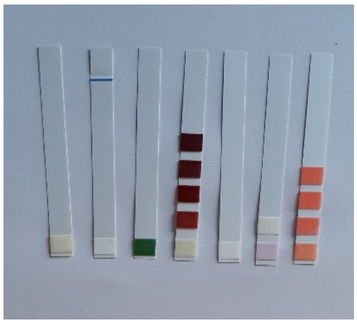
	Ammonium (mg/L)	Carbonate (mg/L)	Chloride (mg/L)	Nitrate (mg/L)
Balcova	ND	183.39	23,352	2,510
Buca	0.0390	125.97	34,529	0.956
Kemalpasa	ND	201.01	192,244	24,193

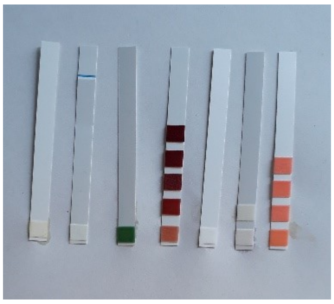
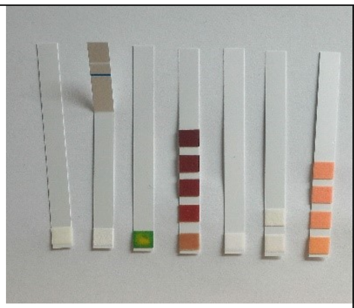
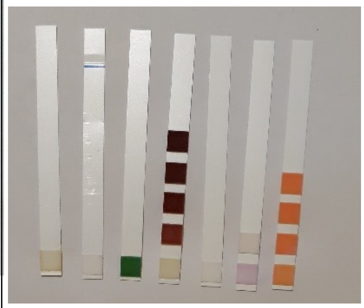
Table S9. Raw data from visual observations

	Ammonium		Carbonate		Chloride		Nitrate	
	Quantity	Concentration (mg/L)	Quantity	Concentration (mg/L)	Quantity	Concentration (mg/L)	Quantity	Concentration (mg/L)
Balcova	2	10	5	285	1	0	2	10
Buca	1	0	4	215	1	0	1	0
Kemalpasa	2	10	5	285	1	0	3	30

Table S10: Water samples were taken from different districts Balcova, Buca and Kemalpasa of Izmir with different illumination types, respectively. The order of ions from left to right is ammonium, arsenic, carbonate, chloride, iron, nitrate and sulfate.

Illumination Types	Balcova	Buca	Kemalpasa
Halogen (H)			
Fluorescent (F)			

Sunlight (S)			
Halogen-Fluorescent (HF)			
Halogen-Sunlight (HS)			
Fluorescent-Sunlight (FS)			
Halogen-Fluorescent-Sunlight (HFS)			

Daylight (D)			

FIGURES

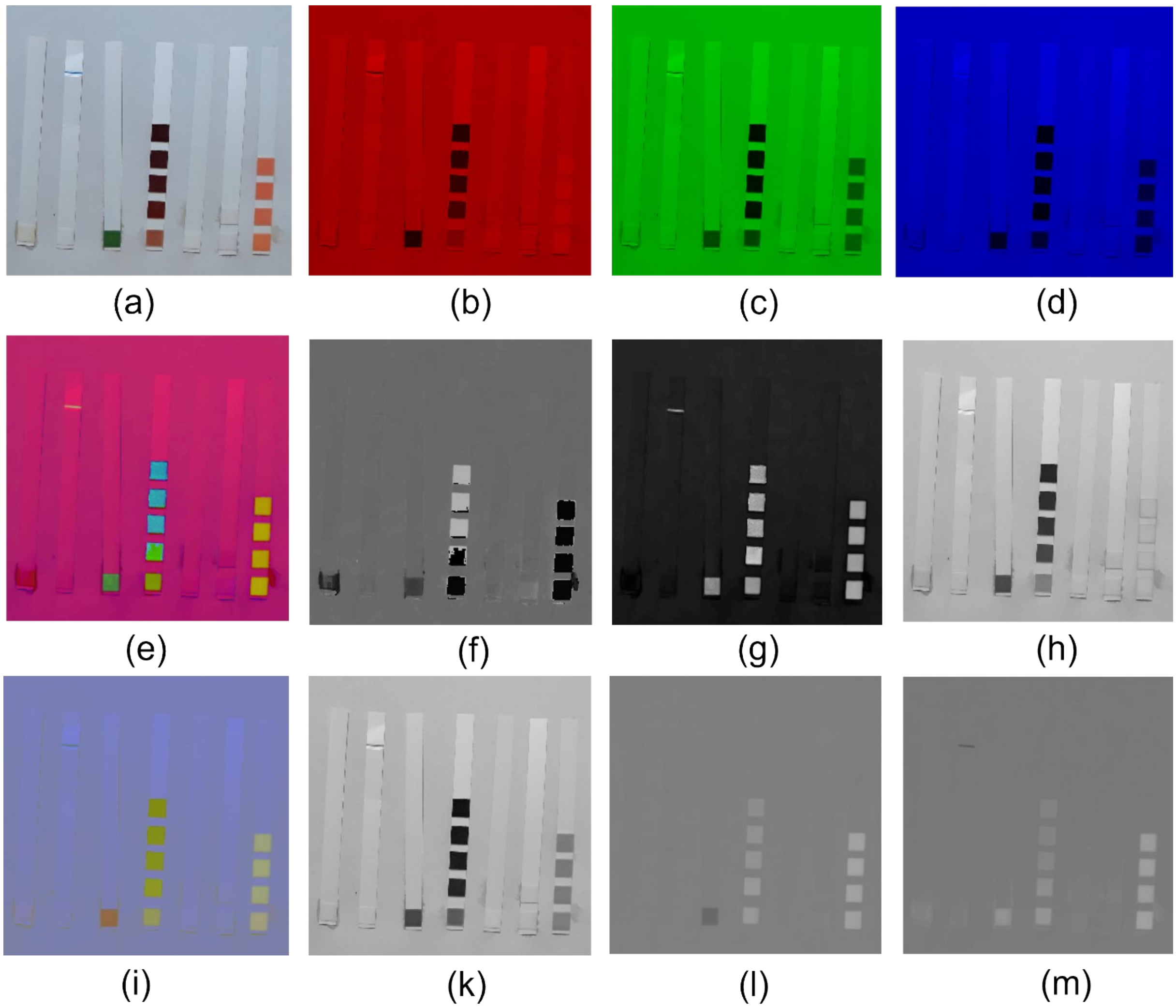


Fig. S1: The original image with RGB in (a), red channel in (b), green channel in (c), blue channel in (d), converted image of RGB to HSV in (e), hue channel in (f), saturation channel in (g), value channel in (h), converted image of RGB to LAB in (i), lightness channel in (k), green-red channel in (l), and blue-yellow channel in (m) are shown.

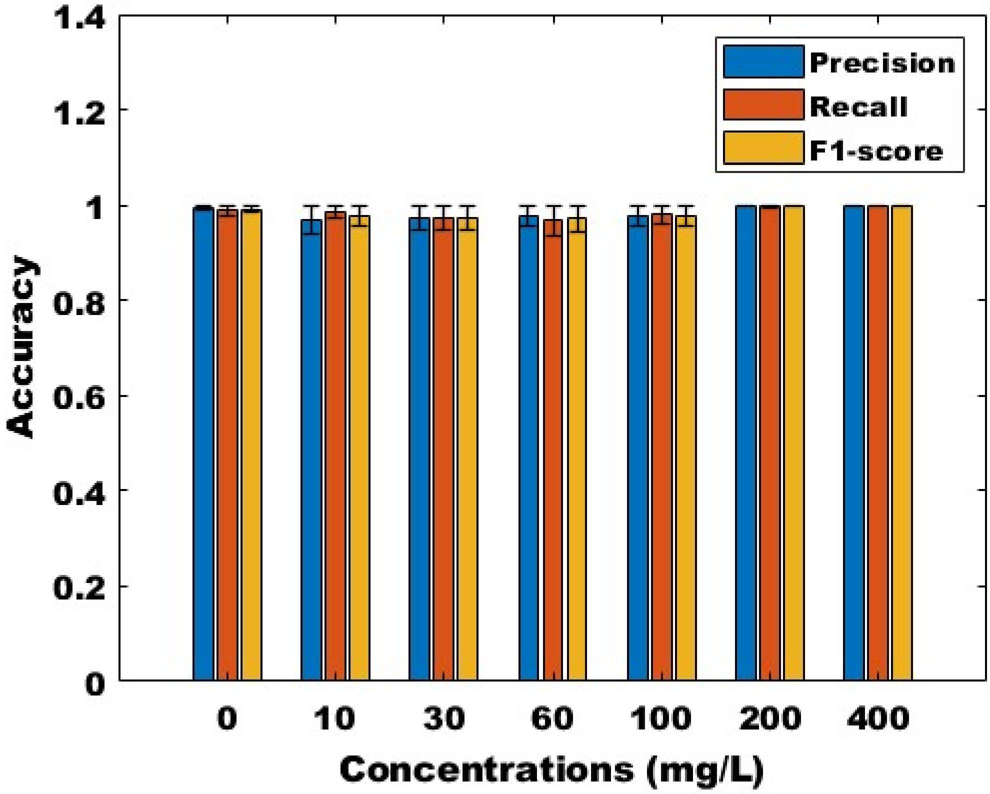


Fig. S2: Evaluation of KNN with error bars in terms of precision, recall, and F1-score for Ammonium.

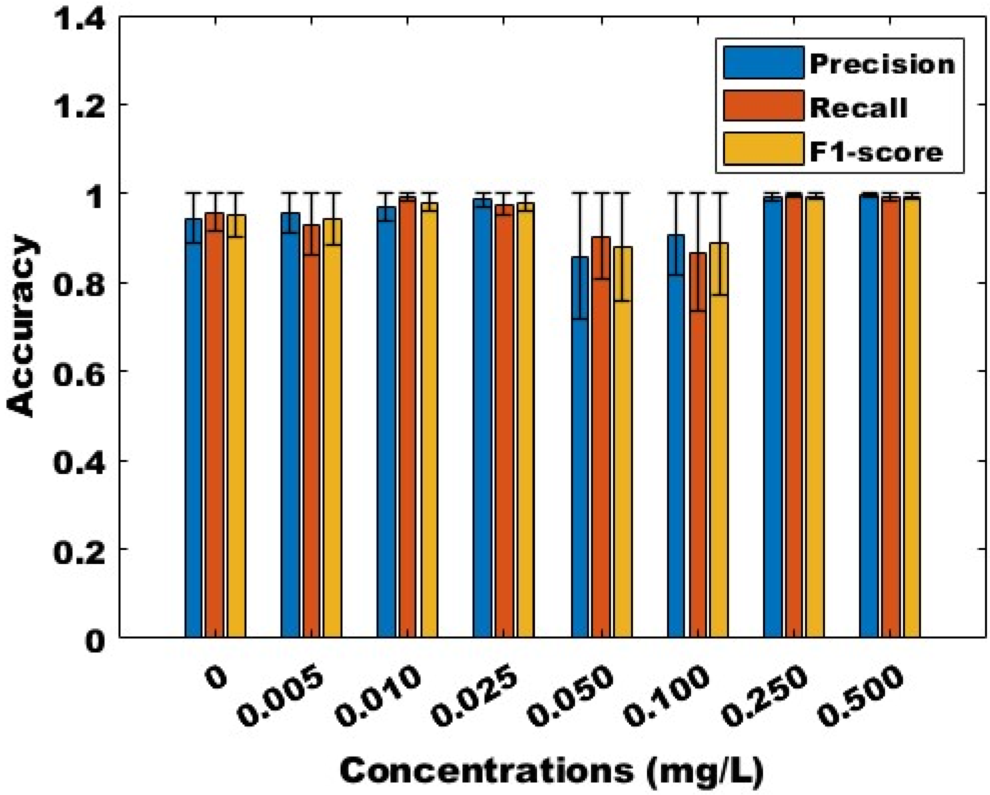


Fig. S3 Evaluation of KNN with error bars in terms of precision, recall, and F1-score for Arsenic.

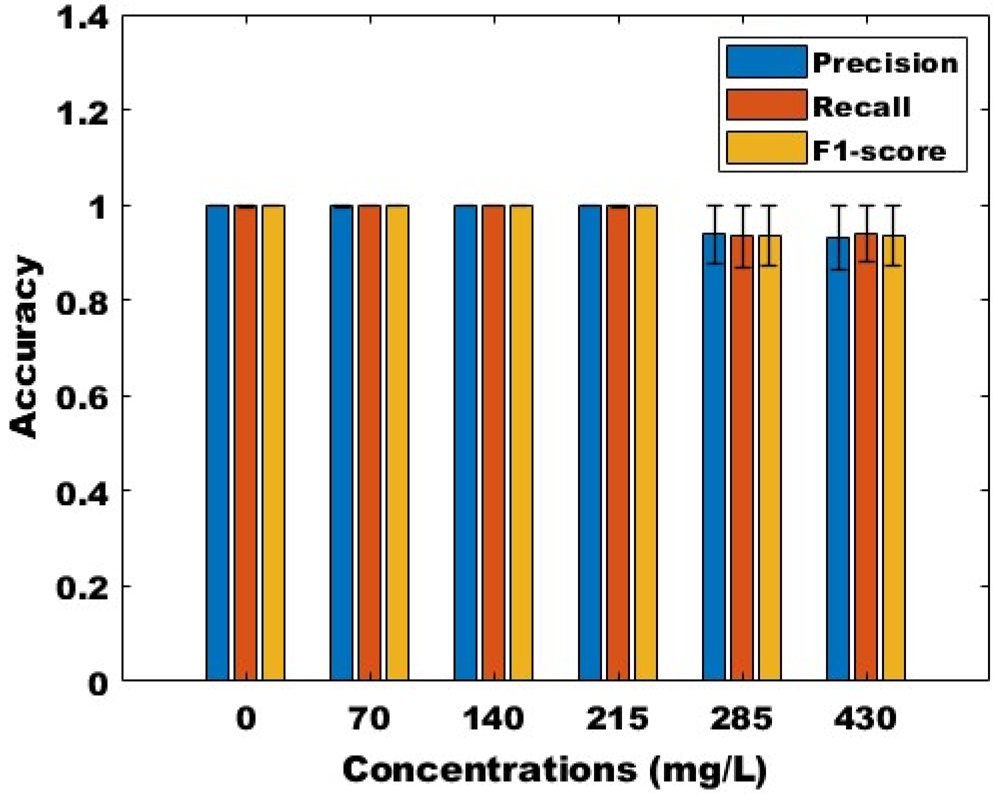


Fig. S4: Evaluation of KNN with error bars in terms of precision, recall, and F1-score for Carbonate.

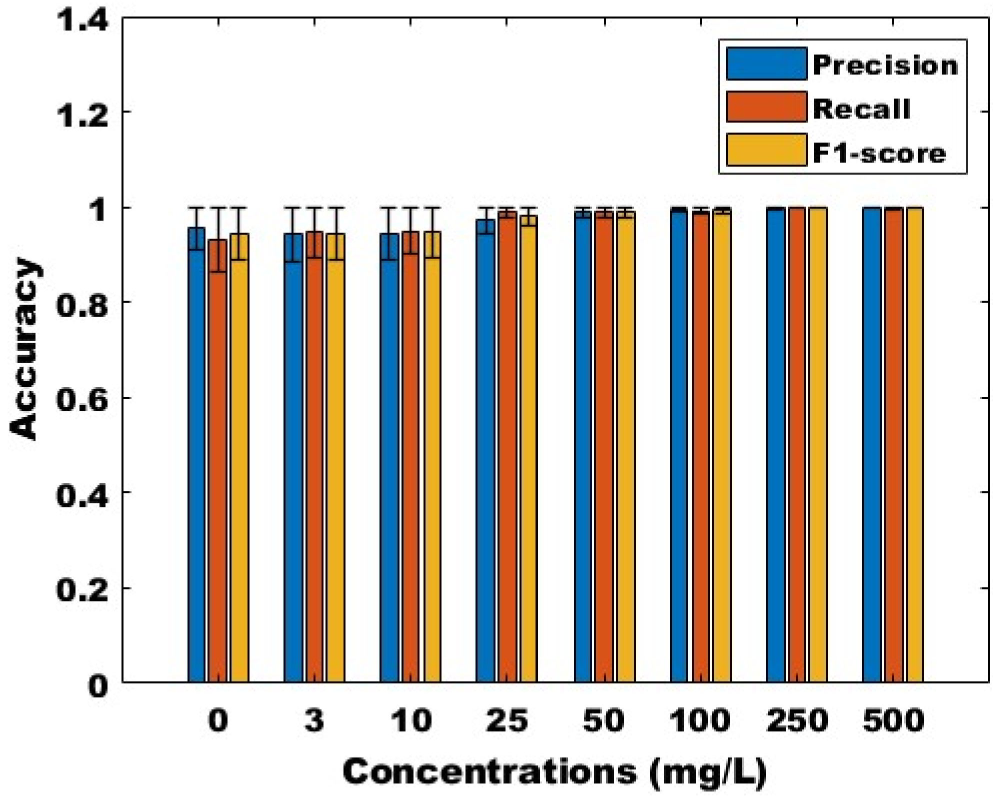


Fig. S5: Evaluation of KNN with error bars in terms of precision, recall, and F1-score for Iron.

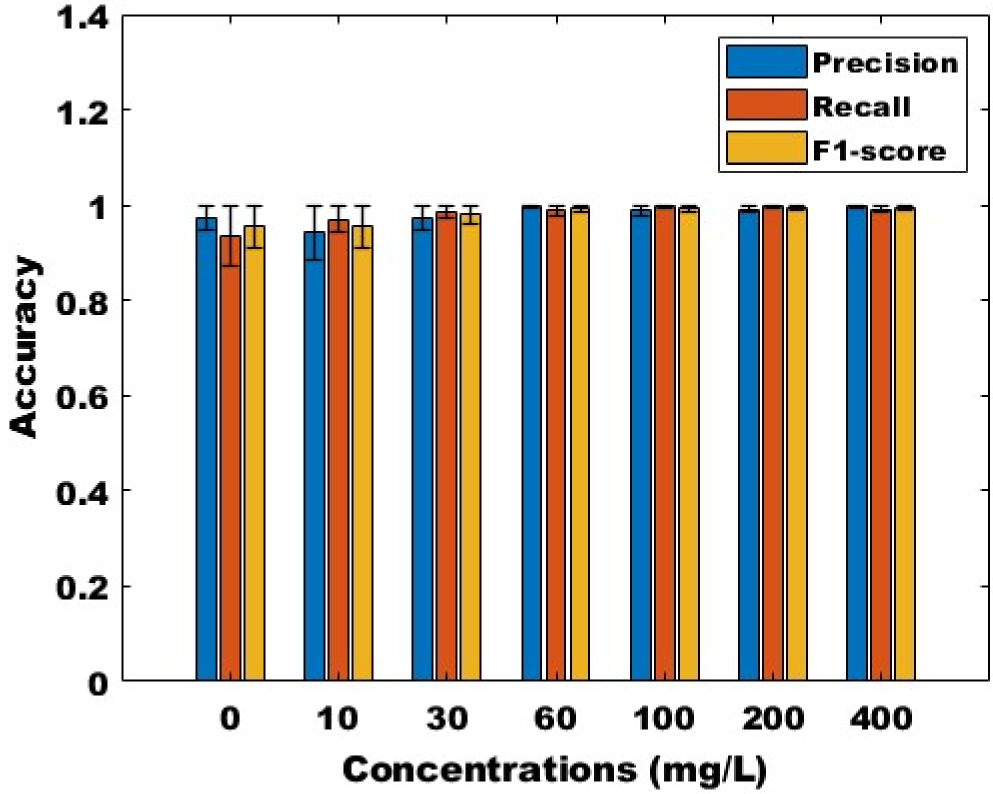


Fig. S6: Evaluation of KNN with error bars in terms of precision, recall, and F1-score for Nitrate.

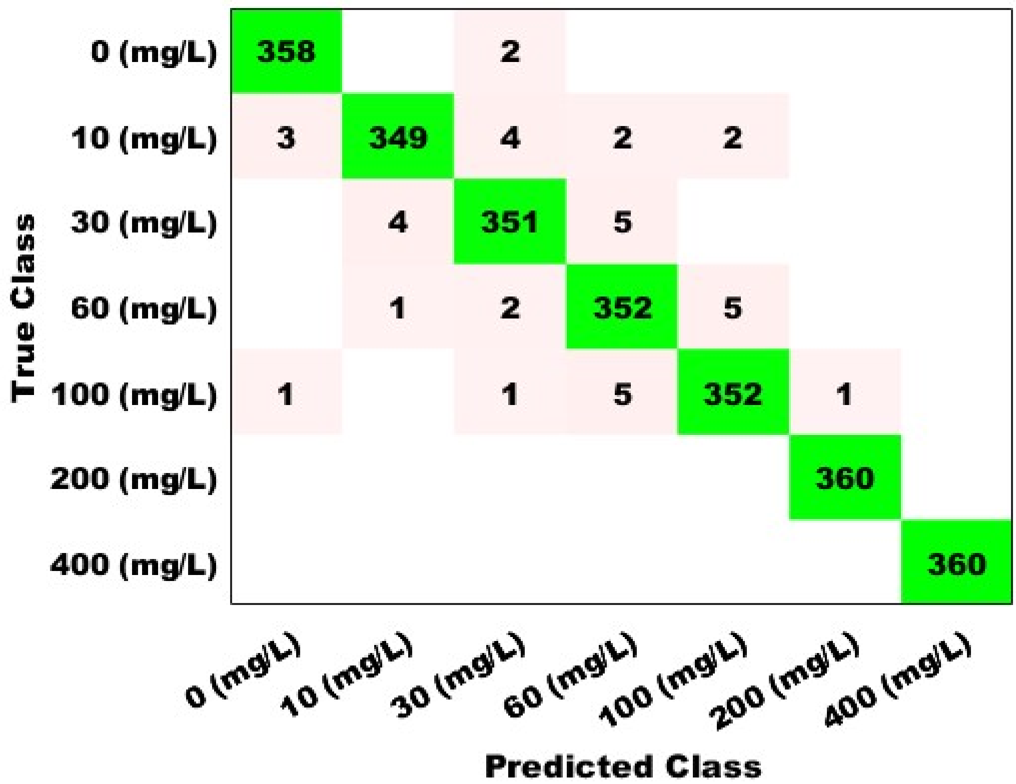


Fig. S7: Confusion matrix of Ammonium for the KNN classifier including 0-400 mg/L concentrations.

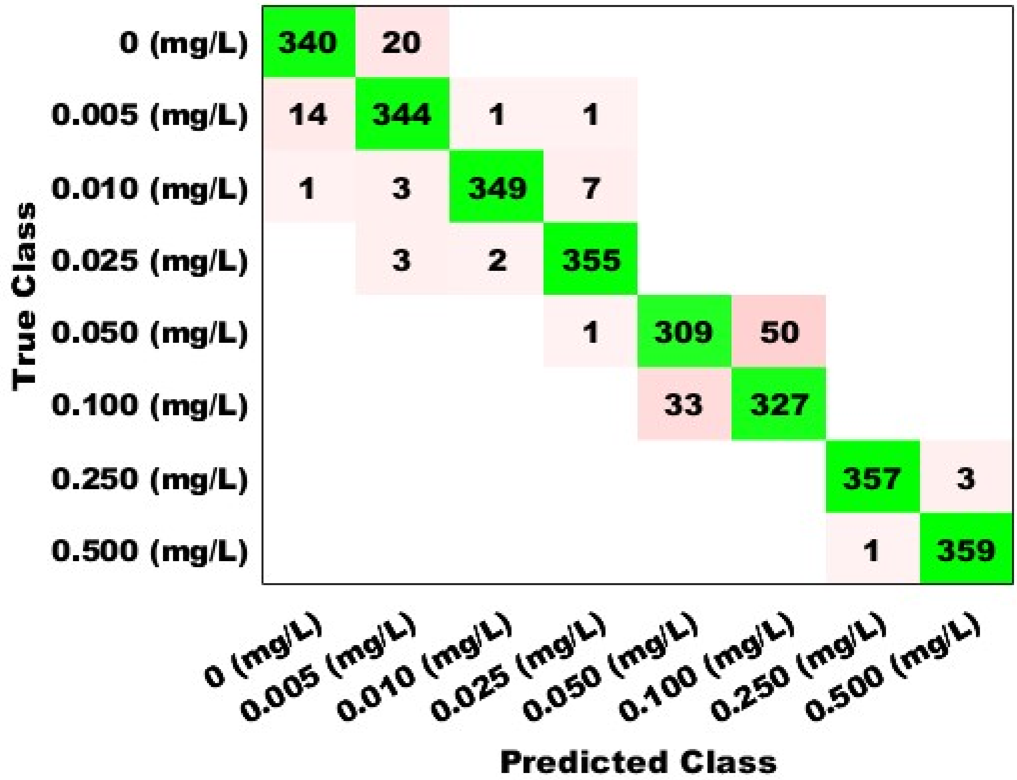


Fig. S8: Confusion matrix of Arsenic for the KNN classifier including 0-0.500 mg/L concentrations.

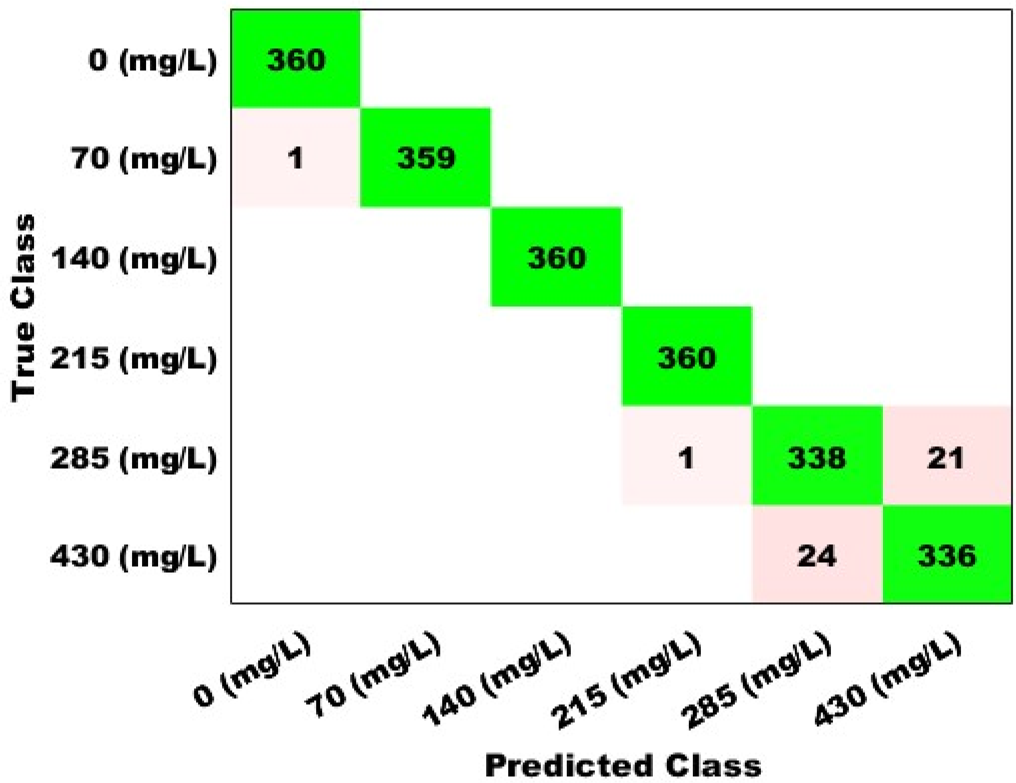


Fig. S9: Confusion matrix of Carbonate for the KNN classifier including 0-430 mg/L concentrations.

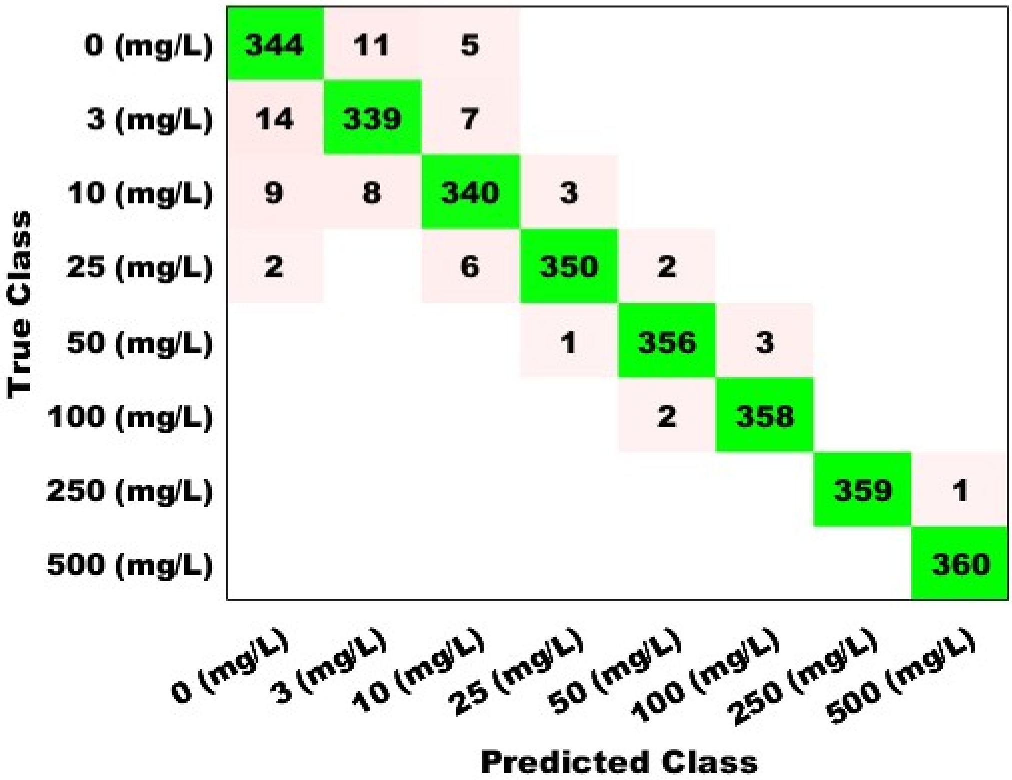


Fig. S10: Confusion matrix of Iron for the KNN classifier including 0-500 mg/L concentrations.

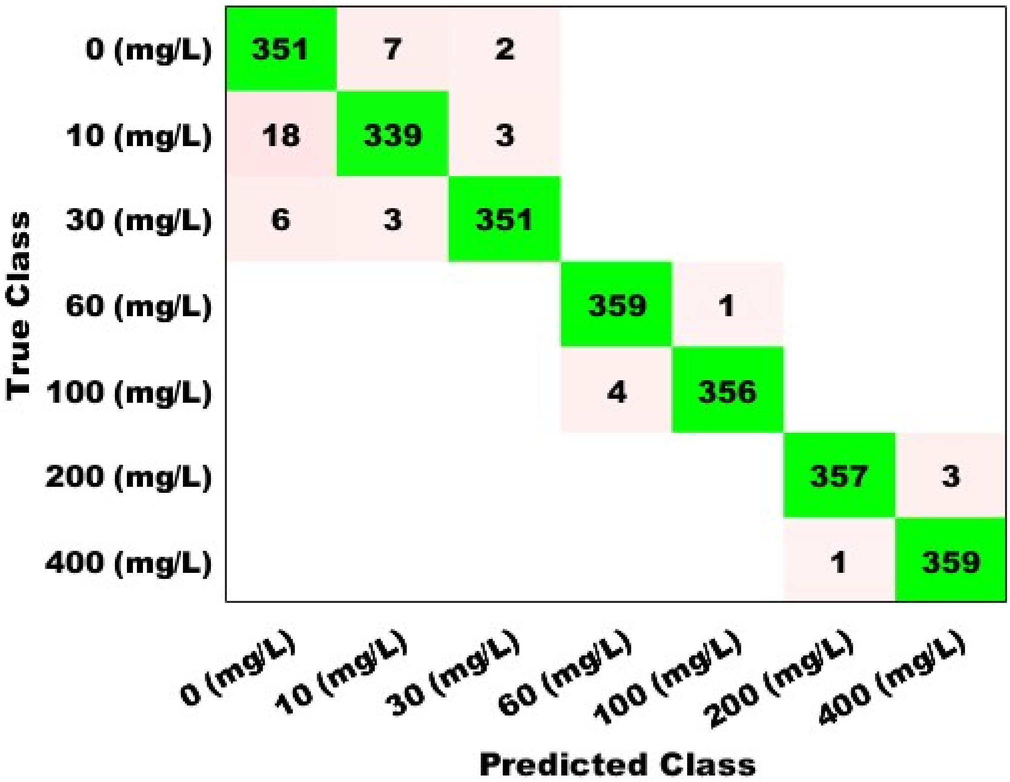


Fig. S11: Confusion matrix of Nitrate for the KNN classifier including 0-400 mg/L concentrations.

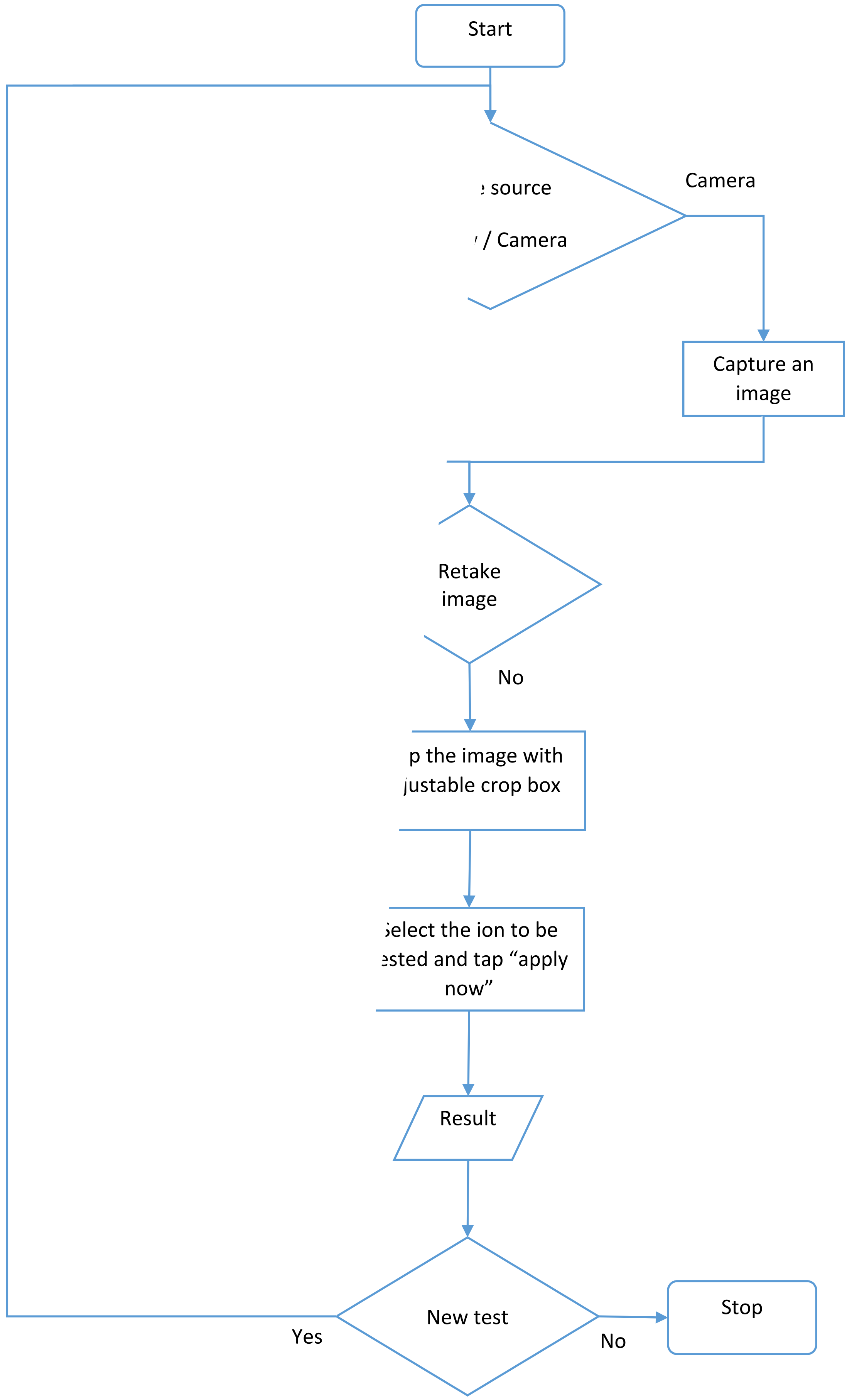


Fig. S12: Flow chart of the proposed water monitoring

S1. Feature extraction

Basic definitions and equations of the features are explained in this section.

1. Mean

The mean, μ, the first color moment, gives the average color value of the image (Equation (1)).

$$\mu = \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q} pixel_{ij} \tag{1}$$

Here, p and q are the values of the pixel in the ith row jth column at image $pixel_{ij}$.

2. Skewness

Skewness is a measure of asymmetry in distribution. The dataset is symmetrical if the left and right sides of the center point are the same. If the skewness is positive (skewed to the right), the data is spread to the left of the mean. If the skewness is negative (skewed to the left), the data will apply to the right of the mean. To extract information from the image, darker and brighter surfaces tend to have positive skewness compared to lighter and matte surfaces. Skewness gives information about the color distribution which is defined in Equation (2),

$$S = \sqrt[3]{\frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}(pixel_{ij} - \mu)^3} \tag{2}$$

3. Kurtosis

Kurtosis is the normalized form of the fourth central moment of distribution. It is also defined as the measure of the sharpness of the peak of a distribution. A high kurtosis distribution usually has a sharper rise, while a low kurtosis distribution usually has a more rounded elevation. Kurtosis is defined in Equation (3),

$$K = \sqrt[4]{\frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}(pixel_{ij} - \mu)^4} \tag{3}$$

4. Contrast

Contrast is a measure of the intensity or gray level variations between a reference pixel and its neighbour. Large contrast indicates large density differences. A still image has a contrast value of 0. The contrast is defined in Equation (4),

$$Contrast = \sum_{i}\sum_{j}(i - j)^2 r(i,j) \tag{4}$$

Here, $r(i,j)$ is the gray level value of the pixel in the (i,j) coordinate.

5. Correlation

Correlation is the measure of linear dependence of gray level values which returns a measure between a pixel and its neighbours. The correlation was calculated in Equation (7) with means μ_i, μ_j (Equation (5)), standard deviations σ_i, σ_j (Equation (6)),

$$\mu_i = \sum_i \sum_j i r(i,j) \quad , \quad \mu_j = \sum_i \sum_j j r(i,j) \tag{5}$$

$$\sigma_i^2 = \sum_i \sum_j (i - \mu_i)^2 r(i,j) \quad , \quad \sigma_j^2 = \sum_i \sum_j (j - \mu_j)^2 r(i,j) \tag{6}$$

$$Correlation = \sum_i \sum_j \frac{(i - \mu_i)(j - \mu_j) r(i,j)}{\sigma_i \sigma_j} \tag{7}$$

6. Homogeneity

Homogeneity is a measure of how close the distribution of elements in the gray-level co-occurrence matrix to the diagonal of the matrix. As homogeneity increases, contrast decreases. Homogeneity is defined in Equation (8),

$$Homogeneity = \sum_i \sum_j \frac{r(i,j)}{1 + |i - j|} \tag{8}$$

7. Energy

The energy property also referred as the angular second-moment property, is a measure of the image homogeneity. It is expressed as the sum of the squares of the matrix elements (Equation (9)),

$$Energy = \sum_i \sum_j r(i,j)^2 \tag{9}$$

8. Entropy

The entropy value was calculated by converting the color input image to a gray level image. The entropy of the image is calculated in Equation (10),

$$Entropy = -\sum n \log_2 n \tag{10}$$

where n is the number of normalized histograms.