Heliyon 5 (2019) e01822

Contents lists available at ScienceDirect

Heliyon

www.elsevier.com/locate/heliyon

A multi-class classification system for continuous water quality monitoring

Swapan Shakhari *, Indrajit Banerjee

Department of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah, West Bengal, 711103, India

ARTICLE INFO

Keywords:
Computer science
Environmental science
Water Quality Monitoring
Evaluation of classification models
Decision tree
C4.5 algorithm
Logistic regression
Data classification
Mathematical analysis
Water resources engineering
Computer science applications
Sensor system
Information systems

ABSTRACT

The issue addressed in this exposition is the classification of multivariate data collected through different sensors for water quality monitoring. Multivariate data are sequences that have various attributes in every instance of the sequences. A few endeavours exist to address this issue; however, none of them has given full emphasis on continuous dataset. Another solution for this issue is to reduce the instances to a single attribute while losing significant information. Different arrangements address both the multivariate and the sequential part of the data yet give an un-versatile solution. The proposed algorithm is not only able to monitor continuous water quality, but it also produces a better classification model for other continuous datasets as well. Instead of decreasing the attributes of the dataset, we introduce three additional reference indicators which are dependent on the actual attributes. We compare the classification accuracy of our proposed algorithm with standard classification models. The proposed method gives better classification accuracy compared to existing methods.

1. Introduction

Water contamination is one of the enormous problems for many decades. Happening green globalisation without sufficient clean water is impossible. The supplied water quality should be recorded and monitored continuously to guarantee the reliable supply of the drinking water. The current state of pollution intensity leads to increased accumulation of pollutants in marine waters at different rates ranging from 7 to 23% depending on the Arctic aquatic environment [1]. In recent times, the problems of remote monitoring systems for detecting and classifying abnormal phenomena on soil or water surface is a great research work [2].

Water influences natural systems [3] and human activities. In 2011, the World Economic Forum (WEF) identified the interconnected resource issues of water, energy, and food as a serious global risk and indicated that managing one aspect of this interrelated system without considering links to the others increases the global threat of serious unintended consequences. Indeed, not managing water and natural resources properly has led to the decline or collapse of civilisations.

1.1. Our contribution

We have developed a real-time water quality monitoring device. The device is consisting of two water quality measurement sensors to collect the primary dataset. An electronic data logging system with a 32-bit ARM core microcontroller board based on the Atmel SAM3X8E ARM Cortex-M3 CPU is used to process the sensors output data and sent these data to a central database system. We have collected water quality data from three different sources; these are “packaged drinking water”, “pipeline drinking water”, and “pond water”.

We propose a decision tree based classification scheme for efficient analysis of the water quality data. We first prepare the training dataset with our collected water quality data. Then, we develop a classification model with this training dataset that can be used to monitor the water quality in real-time automatically. We have classified the primary water quality dataset [4] as well as three secondary water quality datasets available in online repository [5, 6, 7] to test and validate the proposed method.

The rest of the paper is organised as follows. The Section 2 discusses existing related research in water quality monitoring. In Section 3 we have described how the classification method is used for water quality

* Corresponding author.
E-mail address: sshakhari.rs2017@it.iiests.ac.in (S. Shakhari).

https://doi.org/10.1016/j.heliyon.2019.e01822
Received 25 July 2018; Received in revised form 14 April 2019; Accepted 22 May 2019

monitoring. The materials and the proposed method for water quality monitoring is discussed in Section 4. We have presented the comparison of experimental results of our proposed method with C-4.5 Algorithm and Logistic Regression in Section 5. We have also used different comparison metrics that are essential to compare the data classification models in Section 5. And finally, in Section 6 we have concluded the work.

2. Related works

K.M.K. Kut et al. have successfully assessed a total of 176 groundwater samples for its suitability for drinking [8]. Tatjana Mitrović et al. have predicted water quality of Danube River (Serbia) by using Monte Carlo optimised artificial neural networks. They predicted 18 common water quality parameters (WQPs) on inactive monitoring stations [9].

E. Fijani et al. [10] have developed a system for real-time monitoring of two water quality parameters, i.e. chlorophyll-a (Chl-a) and dissolved oxygen (DO). The made a two-layer decomposition using CEEMDAN and VMD algorithms with LSSVM and ELM models.

R. Wan et al. [11] have attempted to develop a four-level pollution index on water quality of marine environment with there classification scheme named Water Quality Classification Index (WQCI).

Multi-Criteria Decision Making Models (MCDM) were adopted by H. Yousefi et al. [12] for evaluating drinking water quality of 190 drinking water wells. The primary objective of the current study was to explore a solution to mitigate probable errors aroused by use of WQI method in the classification of water quality classes.

I.C. Nnorom et al. [13] have studied the aqua physicochemical and potentially toxic elements of ground and surface water sources used for domestic purposes in some districts of Nigeria. They obtained a total of 124 water samples from 13 natural springs, 24 streams, 80 boreholes and seven hand-dug wells were collected from rural and urban areas.

Detailed research work on water quality of Gomti River was done by P. Kumar [14]. He simulated the Water quality along 24 km stretch of the Gomti River from downstream of Near Moosa Bird Sanctuary to Near Bharwara. He had shown the current as well as predicting the future situation using different scenarios while considering critical drivers of global changes namely climate change and population growth.

V. Roth et al. [15] have investigated the effects of climate change on water resources in the transnational Blue Nile Basin (BNB) using water data from the past 25 years. A. Awotwi et al. have shown that the cassava has effects on water yields and water quality components after studying the water quality responses [16].

To our best knowledge, the previous works ware performed using various existing models. In this work, we have presented a specific classification method only for continuous datasets to suit water quality monitoring efficiently.

3. Problem description

The main purpose of this research work is to develop a real-time water quality monitoring system. For the solution of this problem, we had gone through three stages as follows:

- Developing a water quality data logging device that is capable of sending the collected sensor data to our central server (http://wsn.iiests.ac.in:8080).
- Data collection from three different water sources, i.e. 'packaged drinking water', 'pipeline drinking water', and 'pond water' by our water quality data logging device.
- Developing a multi-class classification system for continuous water quality monitoring with the collected water quality data in the central server.

The problem of the multi-class classification system is described below.

Here, we have a continuous-valued training Dataset [D]. Now, we can represent D as a matrix of dimension [R × (N + C)] where, R is the total number of instances, N is the total number of attributes, C = 1, representing the class labels. For example, the following matrix can be used to represent a Dataset [D] with R instances and N attributes.

$$[D] = \begin{bmatrix} r_{0,0} & r_{0,1} & r_{0,2} & \cdots & r_{0,N-1} & c_0 \\ r_{1,0} & r_{1,1} & r_{1,2} & \cdots & r_{1,N-1} & c_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{R-1,0} & r_{R-1,1} & r_{R-1,2} & \cdots & r_{R-1,N-1} & c_{R-1} \end{bmatrix}$$

Here,

$r_{i,j}$ is the jth attribute value of ith instance and c_i is the class label of ith instance where, $0 \le i < R$ & $0 \le j < N$.

In the training Dataset [D] every instance r is given a class label c. But the test dataset will not include the class labels. Our primary objective is to produce a classification model that can be used to determine the class labels of a test dataset more accurately.

4. Materials and methods

4.1. Automatic water quality data logging system

We have developed a water quality data logging device (Fig. 1) consisting of two industrial type water quality measurement sensors pH (pouvoir hydrogen), and TDS (Total Dissolved Solids). We have mounted the pH and TDS sensors in a water pipeline and enclosed into a device to make it portable. The pH sensor can measure 0 to 14 pH with an accuracy of ±0.01 pH. The TDS sensor can weigh 0 to 2000

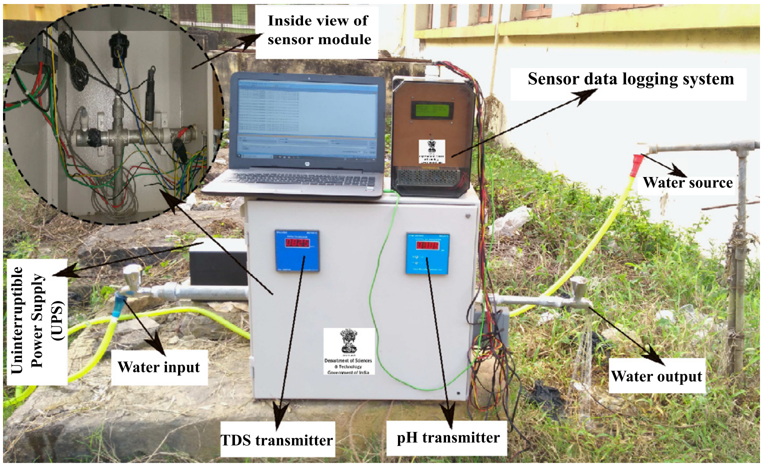

Inside view of sensor module
Sensor data logging system
Water source
Uninterruptible Power Supply (UPS)
Water input
Water output
TDS transmitter
pH transmitter

Fig. 1. A typical water quality data collection point by the monitoring device at location 22°33′17.7″N 88°18′29.6″E (IIEST, Shibpur, India).

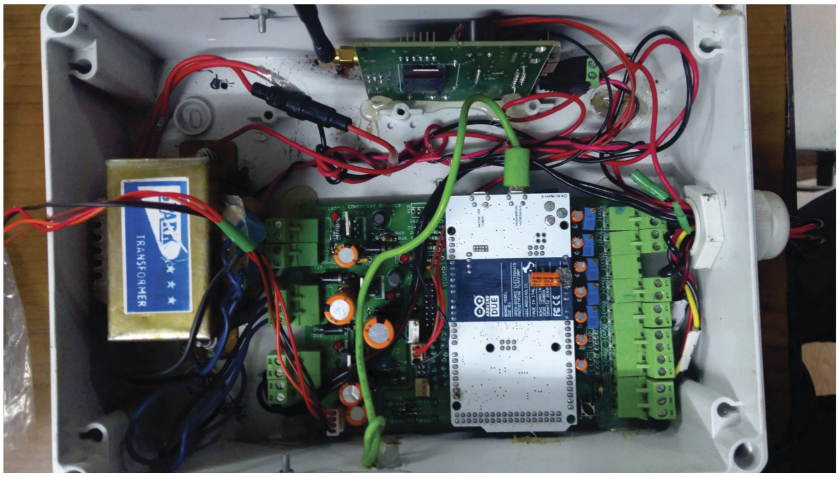


Fig. 2. The inner view of the electronic data logging system with a 32-bit ARM core microcontroller board based on the Atmel SAM3X8E ARM Cortex-M3 CPU.

Table 1
Information of the Datasets which ware classified by the proposed method, C-4.5 [17] & Logistic Regression [18].

DATA SET	#Instances	#Attributes	#Classes
Troxler T., D. Childers. 2018 Dataset [5]	422	8	3
Rosenblatt A. 2014 Dataset [6]	44012	7	2
Steinbach M. 2017 Dataset [7]	486	9	8
Primary Water_Quality_Data [4]	7871	2	3

parts per million (ppm) with ±1% measuring accuracy. The water data signals from the sensors are fetched and converted to numeric values by an Arduino due microcontroller based system (Fig. 2) and logged via serial communication on our data storage and the computing unit.

4.2. *Data collection & preprocessing*

We have collected a total of 7871 instances of water quality data by our water quality monitoring device. But, in our collected primary water quality data, only two water quality parameters are present namely pH and TDS. So, to validate our proposed method for water quality monitoring system, we have also used three additional secondary water quality datasets ([5, 6, 7]). The information about experimented datasets is given in Table 1. We have used Python programming language to access and convert all these datasets as comma-separated value (CSV) format.

We have applied our proposed classification method for water quality monitoring (Algorithm 1) in the datasets as mentioned earlier in Table 1. We also applied two standard methods i.e. C-4.5 [17] and Logistic Regression [18] to classify these datasets (Table 1). We then compared the results of our proposed classification method with these standard classification methods, i.e. C-4.5 and Logistic Regression. We observed a better classification accuracy by our proposed method, thus producing a better water quality monitoring system.

4.3. *Proposed classification method for continuous water quality monitoring*

We have used the tree data structure in our proposed Reference Indicator based Decision Tree algorithm-RIDT [Algorithm 1] for the classification of multi-attribute continuous data. The algorithm is also suited for any integer-valued dataset.

4.3.1. *Introduction of reference indicators*

In our algorithm, we have used three reference indicators (Line No. 4 to 14 of Algorithm 1) with the attributes of the dataset. These additional reference indicators are dependent on all the actual attributes. So, if we make any decision depending on these reference indicators, that decision will be based on all the real attributes together. As we have used these additional reference indicators as well as the original attributes, the essential attributes are not ignored while the least important attributes also considered [Algorithm 1]. As a result, we observed a noteworthy increment for correctly classified instances.

We have used standard deviation and summation of all the attribute values for the first two reference indicators respectively. The first reference indicator is calculated as in Equation (1) where, $\{x_1, x_2, \ldots, x_N\}$ are the actual attribute values of the instances, $\bar{x}$ is the average of the actual attribute values, and N is the number of actual attributes in the instances of the dataset.

$$referenceIndicator_1(instance) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{1}$$

We have tried to map all the attributes together of all instances to possible distinct values by making it dependent on the position and value of actual attributes as well as the second reference indicator. We call it the third reference indicator. This third reference indicator enables us to distinguish the different class instances quickly. It is calculated as in equation (2) where, $\{x_1, x_2, \ldots, x_N\}$ are the actual attribute values of the instances and N is the number of actual attributes in the instances of the dataset.

$$referenceIndicator_3(instance) = mod\left(\sum_{i=1}^{N}(x_i * 2^i), \sum_{i=1}^{N} x_i\right) \tag{2}$$

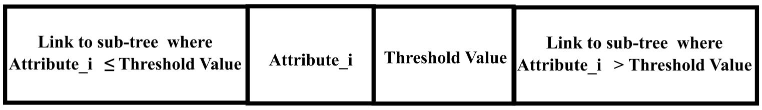
4.3.2. *Selection of the partitioning attribute*

We have used the well-known concept of entropy (E) and information gain (IG) to select the attribute on which the partition will be done. The partition attribute is selected where the maximum information gain is obtained. Every internal node as in Fig. 3 represents a partition of the generated decision tree (Line No. 27 to 33 of Algorithm 1).

The entropy (E) and information gain (IG) calculation in this context are given in Equations (3) and (4) respectively where, X is the training samples, C is the total number of classes, f_i is the fraction of samples which belong to ith class in the training samples, A is the partitioning attribute, X_v is the subset of X having value v for attribute A.

$$E(X) = -\sum_{i=1}^{C} f_i \log_2(f_i) \tag{3}$$

$$IG(X, A) = E(X) - \sum_{v \in Values(A)} \frac{|X_v|}{|A|} E(X_v) \tag{4}$$



Link to sub-tree where Attribute_i ≤ Threshold Value	Attribute_i	Threshold Value	Link to sub-tree where Attribute_i > Threshold Value

Fig. 3. Data structure of every internal node used to hold the classification tree.

Algorithm 1 Proposed Reference Indicator based Decision Tree (RIDT) algorithm for water quality monitoring system.

INPUT: D – A dataset with attributes in CSV format only
OUTPUT: T – A Multivariate continuous data classification tree

```
 1: Tree ← NULL tree node pointer of structure as in Fig. 3
 2: R ← #Instances in D
 3: N ← #Attributes in D
 4: for i ← 0 to R-1 do
 5:     referenceIndicator$_1$ ← Standard Deviation of D[i][0] to D[i][N-2] // last attribute
        (D[i][N-1]) is for class
 6:     referenceIndicator$_2$ ← Sum of D[i][0] to D[i][N-2]
 7:     referenceIndicator$_3$ ← 0.0;
 8:     for j ← 0 to N-2 do
 9:         referenceIndicator$_3$ ← referenceIndicator$_3$ + D[i][j] * power(2,j)
10:     end for
11:     D[i][N] ← referenceIndicator$_1$
12:     D[i][N+1] ← referenceIndicator$_2$
13:     D[i][N+2] ← mod(referenceIndicator$_3$,referenceIndicator$_2$)
14: end for
15: Tree ← DTREE_BUILDER(D)
16: return Tree
17: procedure DTREE_BUILDER(D)
18:     Tree ← NULL tree node pointer of structure as in Fig. 3
19:     R ← #Instances in D
20:     N ← #Attributes in D
21:     if All Instances in D are in one class then
22:         return A leaf node by marking that class
23:     end if
24:     if No attribute is left to make a partition then
25:         return A leaf node marking as the class that is most frequent in these Instances.
26:     end if
        // splitting Attribute selection process starts
27:     maxInfo ← 0.0
28:     for i ← 0 to N-1 do
29:         infoGain ← information gain (IG) if we split D on Attribute$_i$
30:         if maxInfo < infoGain then
31:             maxInfoPossition ← i
32:         end if
33:     end for
        // threshold value of the splitting Attribute calculation starts
34:     Sort the dataset (D) with respect to the values of Attribute$_{maxInfoPossition}$
35:     thresholdValue ← 0.0
36:     for i ← 0 to N-2 do
37:         maxT ← (Attribute$_i$ + Attribute$_{i+1}$)/2
38:         if thresholdValue < maxT then
39:             thresholdValue ← maxT
40:         end if
41:     end for
42:     Tree ← build a tree node that holds Attribute$_{maxInfoPossition}$ and thresholdValue
43:     D$_{v1}$ ← Sub-dataset of D where Attribute$_{maxInfoPossition}$ ⩽ thresholdValue
44:     D$_{v2}$ ← Sub-dataset of D where Attribute$_{maxInfoPossition}$ > thresholdValue
45:     Tree$_{v1}$ ← DTREE_BUILDER(D$_{v1}$)
46:     Attach Tree$_{v1}$ to the left side link of the Tree
47:     Tree$_{v2}$ ← DTREE_BUILDER(D$_{v2}$)
48:     Attach Tree$_{v2}$ to the right side link of the Tree
49:     return Tree
50: end procedure
```

4.3.3. Calculation of threshold value & creation of tree node

The proposed algorithm sets a threshold value and a partition attribute in every internal node. To find the threshold value of a particular partitioning attribute we first sorted the dataset with respect to the values of that partitioning attribute. Then we calculated all the mean values of two consecutive attribute values of the partitioning attribute of this sorted dataset. Finally, we take the maximum mean value as the threshold value (Line No. 34 to 42 of Algorithm 1).

4.3.4. Creation of sub-dataset by splitting the main dataset

After selecting the partitioning attribute and the threshold value, the dataset is divided into two sub-datasets based on this partitioning attribute and the threshold value of that partitioning attribute. One sub-dataset is consists of the instances of the original dataset where the attribute value of the partitioning attribute is less than or equal to the threshold value for that internal node. Rest instances will belong to the other sub-dataset (Line No. 43 and 44 of Algorithm 1).

4.3.5. Creation of leaf node: class determination

In this partitioning process when all the instances of a sub-dataset belong to a single class then, that partition node will be marked as a leaf node with that class label. There may be a situation when making a further partition is not possible because of the unavailability of the partitioning attributes as they have already used in the previous partitioning nodes. Then we marked that partition as a leaf node with the class label that is the most frequent in instances of that sub-dataset partition (Line No. 18 to 26 of Algorithm 1).

At the time of class determination of an instance of the dataset, the threshold values and partitioning attributes of the internal nodes will decide the flow of the comparison path from the root node to the leaf.

4.3.6. Computational time complexity of the proposed algorithm

Lemma 4.1. *The worst case computational complexity to build the classification model is O(RN^2) where R is the total number of instances, and N is the number of attributes in the training dataset.*

Proof. We are splitting the time complexity calculation into the following steps:

Step 1. The inclusion of reference indicators: Computation of the reference indicators for every instance is dependent on the actual number of attributes (N). The algorithm will include the reference indicators for all the instances (R) of the dataset. So, the time complexity for the inclusion of reference indicators is proportional to RN (Line no 4 to 14 of Algorithm 1).

Step 2. Creation of leaf node: The proposed algorithm will depend on the number of instances (R) in the current dataset to check whether leaf node creation criteria is mate or not (Line no 22 and 24 of Algorithm 1). Leaf node creation time is constant (Line no 22 and 25 of Algorithm 1).

Step 3. Selection of the partitioning attribute: For the selection of the partitioning attribute (Line no 27 to 33 of Algorithm 1), the proposed algorithm will calculate the information gains of the current dataset at all the attribute positions, so the maximum time will depend on the number of attributes (N) of the current dataset.

Step 4. Dataset sorting and threshold value calculation: In line number 34 to 41 of Algorithm 1, the dataset sorting time and threshold value calculation time for a partitioning attribute will depend on the number of instances of the current dataset (R).

Step 5. Building a tree node and divide the current dataset into two sub-datasets: The tree node building time is constant in Line no 42 of Algorithm 1. The creation of sub-datasets also takes constant time in line no. 43 and 44 of Algorithm 1.

Step 6. Recursive call of the DTREE_BUILDER procedure: The proposed algorithm will recursively call the DTREE_BUILDER procedure in line no 45 and 47 of Algorithm 1.

In the worst case scenario, the proposed algorithm will build a complete tree where every path test every attribute. So, the total levels of the tree will be N. All the instances (R) will spread across all the nodes at each of the N levels.

At each level i, in the tree the algorithm must examine the remaining N – i attributes for each instance at the level to calculate the information gains and threshold values. So, the maximum number of time for recursively calling the DTREE_BUILDER procedure is the number of attributes (N) in the original dataset.

So, the overall computational complexity to produce the classification model by the proposed algorithm will be O(RN + R + N + R + constant time) × N ≈ O(RN × N) ≈ O(RN^2). □

However, in most of the cases, the number of leaves are ≪ R. In practice, complexity is linear in both number of attributes (N) and the number of training instances (R).

5. Results and discussion

5.1. Evaluation of proposed method by confusion matrix

A confusion matrix demonstrates the quantity of right and wrong estimates produced by the classification model. It shows contrast with the genuine results (original class) in the dataset. The matrix dimension is N × N. N is the quantity of original classes. Utilizing the data in the confusion matrix, we can evaluate the performance of the classification models.

The confusion matrixes generated from the outputs of our proposed method, C-4.5, and Logistic Regression is given in Table 2 for our primary water quality dataset. The confusion matrixes for Troxler T., D. Childers. 2018 water quality dataset [5], Rosenblatt A. 2014 water quality dataset [6] and Steinbach M. 2017 water quality dataset [7] is produced on Tables 3, 4 and 5 respectively.

In our confusion matrices in the Tables 2, 3, 4 and 5, the row values represent the actual attribute classes of the dataset and the column values represent the predicted classes by the classification model. The total instances of an actual class are the summation of the values of the corresponding row. The total instances of a certain class that are predicted by the classification model are the summation of the corresponding column values. The principal diagonal of the confusion matrix represents the correctly classified instances of the corresponding class of the row, column position.

5.1.1. Different prediction cases of the proposed method

There are four cases when a classifier model makes predictions. These are as follows:

N_{TP} (Number of True Positives): These are cases in which the classifier model predicted some instances to a certain class, and they actually do belongs to that predicted class.

Table 2
Confusion matrix of primary water dataset generated from the output of Proposed Method, C-4.5 [17] & Logistic Regression [18].

By Proposed Method and C-4.5 Algorithm			
Water source	Packaged drinking water	Pipeline drinking water	Pond water
Packaged drinking water	**2689**	0	0
Pipeline drinking water	0	**2627**	0
Pond water	0	0	**2555**

By Logistic Regression			
Water source	Packaged drinking water	Pipeline drinking water	Pond water
Packaged drinking water	**2672**	17	0
Pipeline drinking water	59	**2202**	366
Pond water	57	293	**2205**

Table 3
Confusion matrix of Troxler T., D. Childers. 2018 water quality dataset [5] generated from the output of Proposed Method, C-4.5 [17] & Logistic Regression [18]. Here, TS/PH6a, TS/PH8, and TS/PH7a are three water quality data logging sites having Longitude (degree), Latitude (degree) at −80.649, 25.214; −80.525, 25.233, and −80.639, 25.191 respectively.

By Proposed Method			
Site name	TS/PH6a	TS/PH8	TS/PH7a
TS/PH6a	**145**	8	30
TS/PH8	17	**32**	5
TS/PH7a	12	1	**172**

By C-4.5 Algorithm			
Site name	TS/PH6a	TS/PH8	TS/PH7a
TS/PH6a	**152**	1	30
TS/PH8	27	**23**	4
TS/PH7a	23	3	**159**

By Logistic Regression			
Site name	TS/PH6a	TS/PH8	TS/PH7a
TS/PH6a	**128**	3	52
TS/PH8	32	**1**	21
TS/PH7a	74	3	**108**

Table 4
Confusion matrix of Rosenblatt A. 2014 water quality dataset [6] generated from the output of Proposed Method, C-4.5 [17] & Logistic Regression [18].

By Proposed Method		
Bay/River	Tarpon Bay	Shark River
Tarpon Bay	**21087**	945
Shark River	1126	**20854**

By C-4.5 Algorithm		
Bay/River	Tarpon Bay	Shark River
Tarpon Bay	**21103**	929
Shark River	1563	**20417**

By Logistic Regression		
Bay/River	Tarpon Bay	Shark River
Tarpon Bay	**17994**	4038
Shark River	4456	**17524**

N_{TN} (Number of True Negatives): These are cases in which the classifier model predicted some instances, not to belong to a certain class, and they actually do not belong to that predicted class.

N_{FP} (Number of False Positives): These are cases in which the classifier model predicted some instances to a certain class, but they actually do not belong to that predicted class.

N_{FN} (Number of False Negatives): These are cases in which the classifier model predicted some instances, not to belong to a certain class, but they actually do belong to that predicted class.

We can calculate the values of these cases from the confusion matrix. Number of instances in the test dataset is the summation of $\mathbf{N_{TP}}$, $\mathbf{N_{TN}}$, $\mathbf{N_{FP}}$ and $\mathbf{N_{FN}}$.

5.1.2. Evaluation metrics of the Proposed Method for each individual classes

With the help of the four prediction cases of the classification models described in Section 5.1.1, we can formulate some evaluation metrics [19, 20], to observe it's performance on each individual class, such as Sensitivity (Sens) or True Positive Rate (TPR), False Positive Rate (FPR), Specificity (Spec) or TNR, Positive Predictive Value (PPV) or Precision, Negative Predictive Value (NPV), Accuracy (Acc), F1 score,

Table 5
Confusion matrix of Steinbach M. 2017 water quality dataset [7] generated from the output of Proposed Method, C-4.5 [17] & Logistic Regression [18]. Here, BL = BUG LAKE, BLD = BUG LAKE DEEP, CL = CLOUD LAKE, CLD = CLOUD LAKE DEEP, DL = DEVILS LAKE, DLD = DEVILS LAKE DEEP, KL = KING LAKE, & KLD = KING LAKE DEEP.

By Proposed Method

Class	BL	BLD	CL	CLD	DL	DLD	KL	KLD
BL	**74**	1	0	0	1	0	0	0
BLD	10	**32**	0	0	0	0	0	0
CL	1	3	**72**	5	0	0	0	0
CLD	0	0	4	**38**	0	0	0	1
DL	0	0	0	0	**74**	3	0	0
DLD	1	2	0	0	6	**34**	0	0
KL	1	0	0	0	0	0	**79**	1
KLD	0	0	0	0	0	0	5	**38**

By C-4.5 Algorithm

Class	BL	BLD	CL	CLD	DL	DLD	KL	KLD
BL	**72**	2	0	0	1	1	0	0
BLD	4	**37**	0	0	0	1	0	0
CL	1	3	**71**	6	0	0	0	0
CLD	0	0	4	**38**	0	0	0	1
DL	0	0	0	0	**72**	5	0	0
DLD	1	2	0	0	7	**33**	0	0
KL	1	0	0	0	0	0	**79**	1
KLD	0	0	0	0	0	0	5	**38**

By Logistic Regression

Class	BL	BLD	CL	CLD	DL	DLD	KL	KLD
BL	**72**	2	0	0	1	1	0	0
BLD	13	**26**	0	1	1	1	0	0
CL	2	0	**71**	5	0	0	3	0
CLD	0	0	18	**25**	0	0	0	0
DL	6	0	0	0	**66**	5	0	0
DLD	0	2	0	0	17	**24**	0	0
KL	1	1	0	0	0	0	**77**	2
KLD	0	0	0	0	0	0	4	**39**

and Matthews Correlation Coefficient (MCC) as defined in Equations (5), (6), (7), (8), (9), (10), (11) and (12) respectively.

$$\mathrm{TPR} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} \tag{5}$$

$$\mathrm{FPR} = \frac{N_{\mathrm{FP}}}{N_{\mathrm{TN}} + N_{\mathrm{FP}}} \tag{6}$$

$$\mathrm{Spec} = \frac{N_{\mathrm{TN}}}{N_{\mathrm{TN}} + N_{\mathrm{FP}}} \tag{7}$$

$$\mathrm{PPV} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}} \tag{8}$$

$$\mathrm{NPV} = \frac{N_{\mathrm{TN}}}{N_{\mathrm{TN}} + N_{\mathrm{FN}}} \tag{9}$$

$$\mathrm{Acc} = \frac{N_{\mathrm{TP}} + N_{\mathrm{TN}}}{N_{\mathrm{TP}} + N_{\mathrm{TN}} + N_{\mathrm{FP}} + N_{\mathrm{FN}}} \tag{10}$$

$$\text{F1 score} = \frac{2N_{\mathrm{TP}}}{2N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{FN}}} \tag{11}$$

$$\mathrm{MCC} = \frac{N_{\mathrm{TP}} \times N_{\mathrm{TN}} - N_{\mathrm{FP}} \times N_{\mathrm{FN}}}{\sqrt{(N_{\mathrm{TP}} + N_{\mathrm{FP}})(N_{\mathrm{TP}} + N_{\mathrm{FN}})(N_{\mathrm{TN}} + N_{\mathrm{FP}})(N_{\mathrm{TN}} + N_{\mathrm{FN}})}} \tag{12}$$

In Fig. 4 and 5, the Receiver Operating Characteristic (ROC) curve [21] is produced for our primary water quality dataset and Troxler T., D. Childers. 2018 water quality dataset [5] by plotting the true positive rate (Equation (5)) against the false positive rate (Equation (6)) to visualize the performance of the proposed method. The ROC curves for water quality datasets of Rosenblatt A. 2014 [6] and Steinbach M. 2017 [7] are given in Fig. 6 and 7 respectively.

Table 6
Percentage of Correctly Classified Instances generated from the output of C-4.5 algorithm [17], logistic regression [18] & our classification algorithm.

DATA SET	C-4.5	Logistic regression	Proposed method
Troxler T., D. Childers. 2018 Dataset [5]	79.15%	56.16%	82.70%
Rosenblatt A. 2014 Dataset [6]	94.34%	80.70%	95.29%
Steinbach M. 2017 Dataset [7]	90.54%	82.30%	90.74%
Primary Water_Quality_Data [4]	100.00%	89.94%	100.00%

5.1.3. Overall evaluation metrics of the classification models

Definition 1. The **Factor of Correctly Classified Instances (FCCI)** of a classification model can be calculated as the ratio of the summation of Number of True Positives of every class with the total number of instances present in the dataset.

The Equation (13) defines the calculation of FCCI where C is the number of classes in the dataset and $N_{\mathrm{TP}}{}^{i}$ is the number of true positives of class i. We can obtain the Percentage of Correctly Classified Instances (PCCI) by multiplying FCCI with 100, i.e. PCCI = FCCI × 100. In Table 6, we have shown the Percentage of Correctly Classified Instances (PCCI) of six datasets by C-45, Logistic Regression, and the proposed algorithm.

We observed that the Percentage of Correctly Classified Instances (PCCI) of the proposed algorithm is more considerable than C-4.5 and Logistic Regression for four datasets. For the water dataset, we ware able to achieve PCCI = 100% (Table 6) with the proposed algorithm and C-4.5.

$$\mathrm{FCCI} = \frac{\sum_{i=1}^{C} N_{\mathrm{TP}}{}^{i}}{\text{Number of Instances in the Dataset}} \tag{13}$$

C-4.5 [17] & Logistic Regression [18]

Definition 2. The **Mean Absolute Error (MAE)** is a linear score which implies that all the individual differences (errors) are given the same weight by measuring the average of the errors in an arrangement of forecasts. It also shows the closeness of estimates to the actual.

The MAE is represented by the equation (14) where f_i = prediction and y_i = true value and average of the absolute error $|e_i| = |f_i - y_i|$.

$$MAE = \frac{1}{n}\Sigma_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\Sigma_{i=1}^{n}|e_i| \tag{14}$$

We have accumulated the Mean Absolute Errors of the six datasets by Logistic Regression, C-4.5, and our proposed algorithm. We show this MEA values in Fig. 8 as a bar graph visualisation. We observed that our algorithm is producing minimum MEA over C-4.5 and Logistic Regression.

Definition 3. Root Mean Squared Error (RMSE) is the squared average difference amongst estimate and relating observed values of different models for a particular dataset and not between datasets. The errors are squared before they are found the middle value.

The RMSE gives high weight to huge errors. This means the RMSE is most helpful when huge errors are not acceptable. The Root Mean Squared Error (RMSE) is represented in equation (15) that measures the contrasts between the values really show in the occurrences and the values estimated by a classification model [22, 23].

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{j=1}^{N}\left(\frac{P_{i,j} - T_j}{T_j}\right)^2} \tag{15}$$

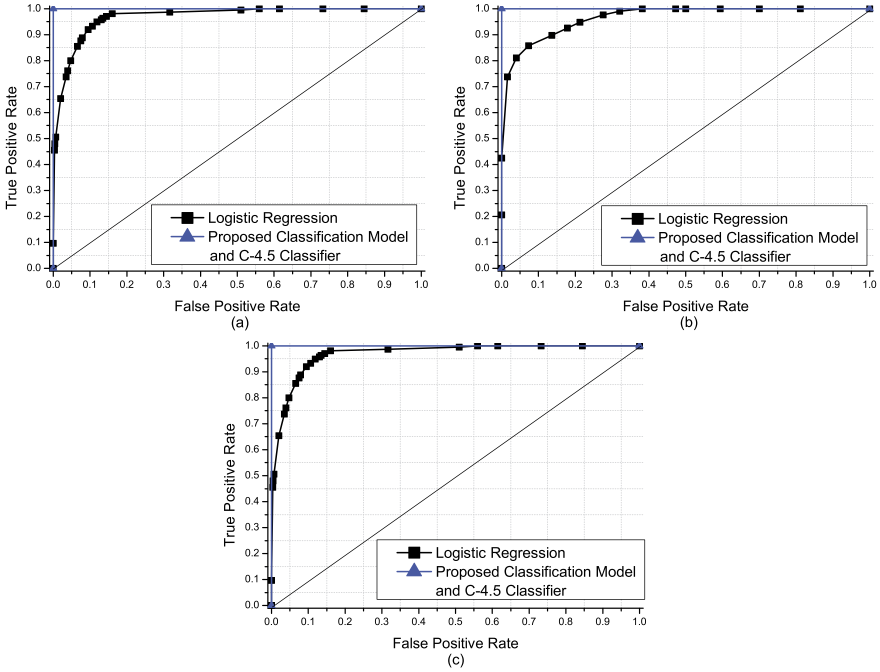


Fig. 4. Receiver operating characteristic (ROC) curve of primary water quality dataset by the proposed method, C-4.5 Classifier & Logistic Regression. Here, (a) ROC curve for "packaged drinking water" quality, (b) ROC curve for "pipeline drinking water" quality and (c) ROC curve for "pond water" quality.

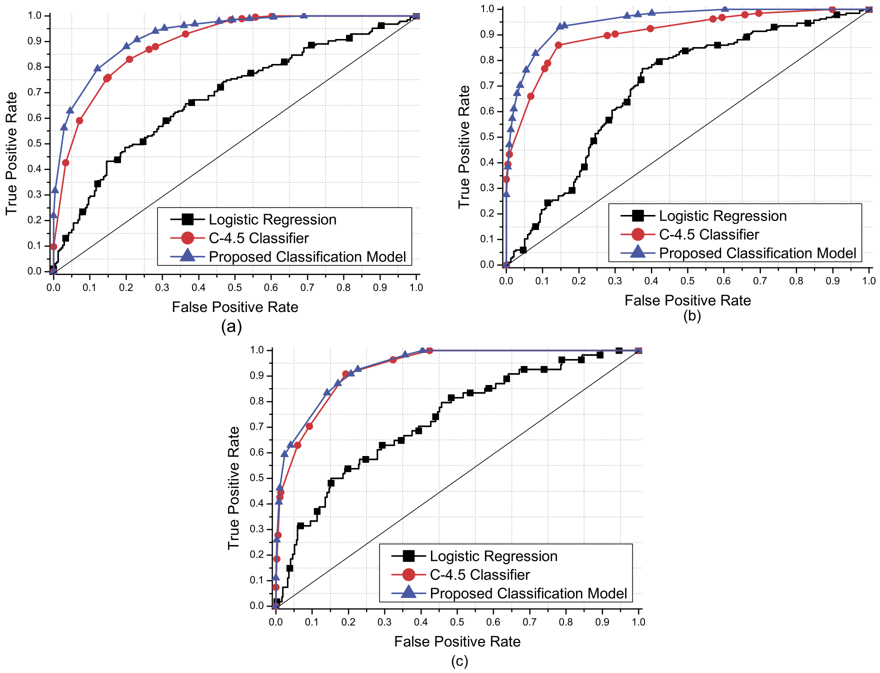


Fig. 5. Receiver operating characteristic (ROC) curve of Troxler T., D. Childers. 2018 water quality dataset [5] by the proposed method, C-4.5 Classifier & Logistic Regression. Here, (a) ROC curve for PH6a water quality having Site Coordinates at Longitude (degree): −80.649 & Latitude (degree): 25.214, (b) ROC curve for PH7a water quality having Site Coordinates at Longitude (degree): −80.639 & Latitude (degree): 25.191, and (c) ROC curve for PH8 water quality having Site Coordinates at Longitude (degree): −80.525 & Latitude (degree): 25.233.

Here, $P_{(i,j)}$ is the predicted value, i fitness case, T_j is the target value of the fitness case j and N is the number of instances.

We have calculated RMSE values by C-4.5, Logistic Regression, and our proposed algorithm and presented a comparison of these RMSE values in Fig. 9. The proposed classification algorithm can maintain a lower Root Mean Squared Error compared to C-4.5 and Logistic Regression.

Definition 4. Kappa Statistic adjusts the level of agreement between the classifier's forecasts and reality by considering the extent of expectations that may happen by chance.

Kappa Statistic in equation (16) and (18) compares the precision of the classification framework to the exactness of an irregular framework, i.e. the results given randomly [24, 25]. A random model can also pro-

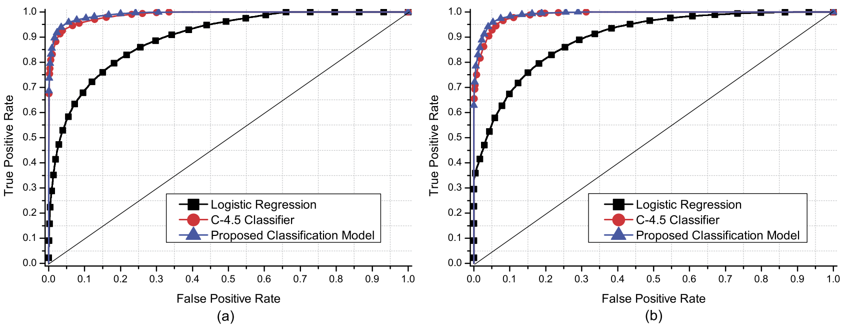


Fig. 6. Receiver operating characteristic (ROC) curve of Rosenblatt A. 2014 water quality dataset [6] by the proposed method, C-4.5 Classier & Logistic Regression. Here, (a) ROC curve of Shark River, (b) ROC curve of Tarpon Bay.

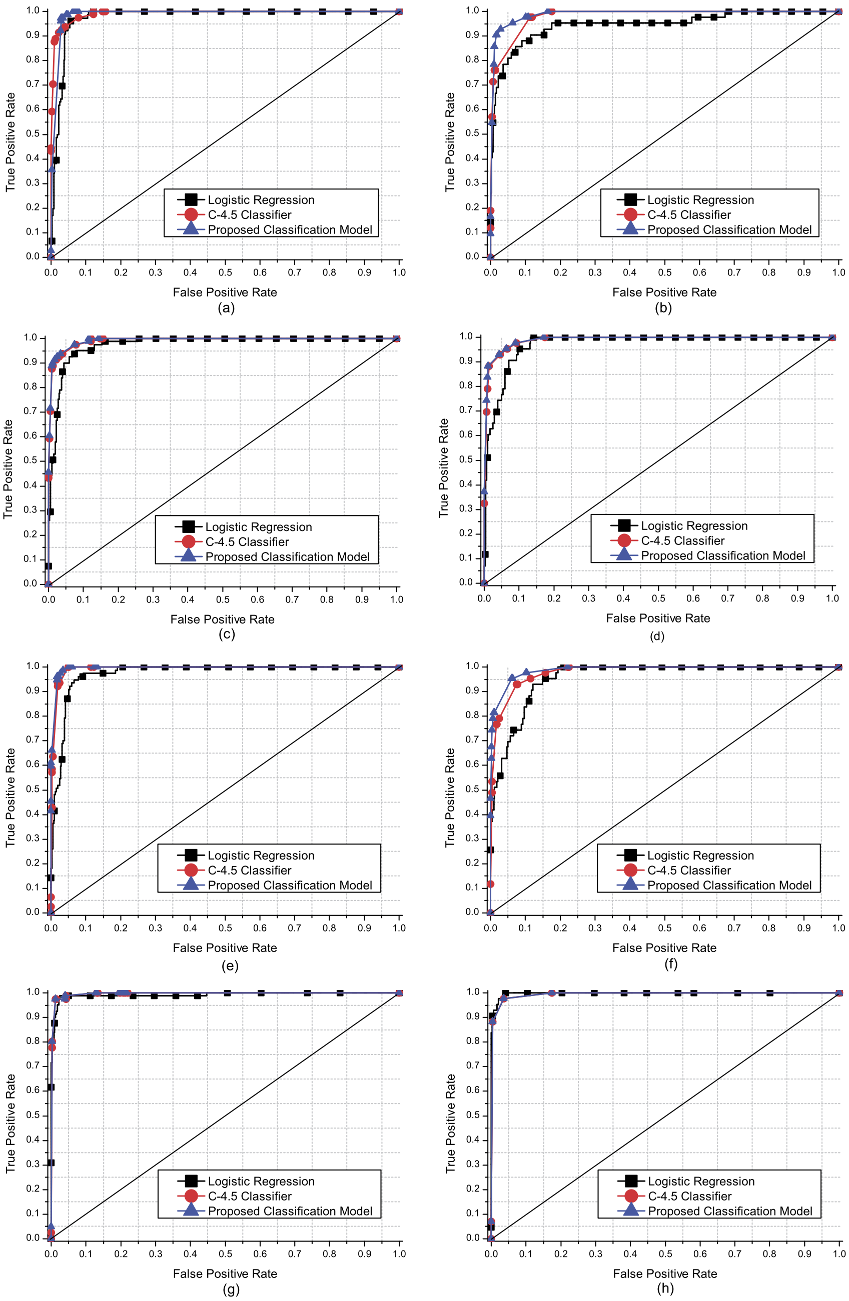


Fig. 7. Receiver operating characteristic (ROC) curve of Steinbach M. 2017 water quality dataset [7] by the proposed method, C-4.5 Classier & Logistic Regression. Here, (a) ROC curve of BUG LAKE, (b) ROC curve of BUG LAKE DEEP, (c) ROC curve of CLOUD LAKE, (d) ROC curve of CLOUD LAKE DEEP, (e) ROC curve of DEVILS LAKE, (f) ROC curve of DEVILS LAKE DEEP, (g) ROC curve of KING LAKE, & (h) ROC curve of KING LAKE DEEP.

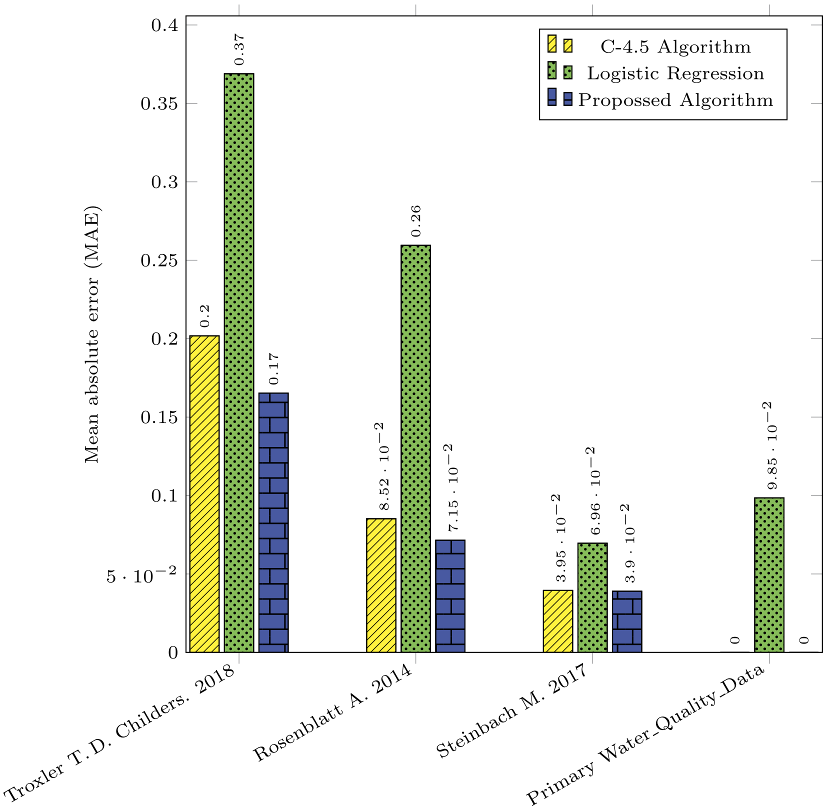


Fig. 8. Mean absolute error of C-4.5 algorithm; logistic regression & proposed algorithm for the dataset of Table 1.

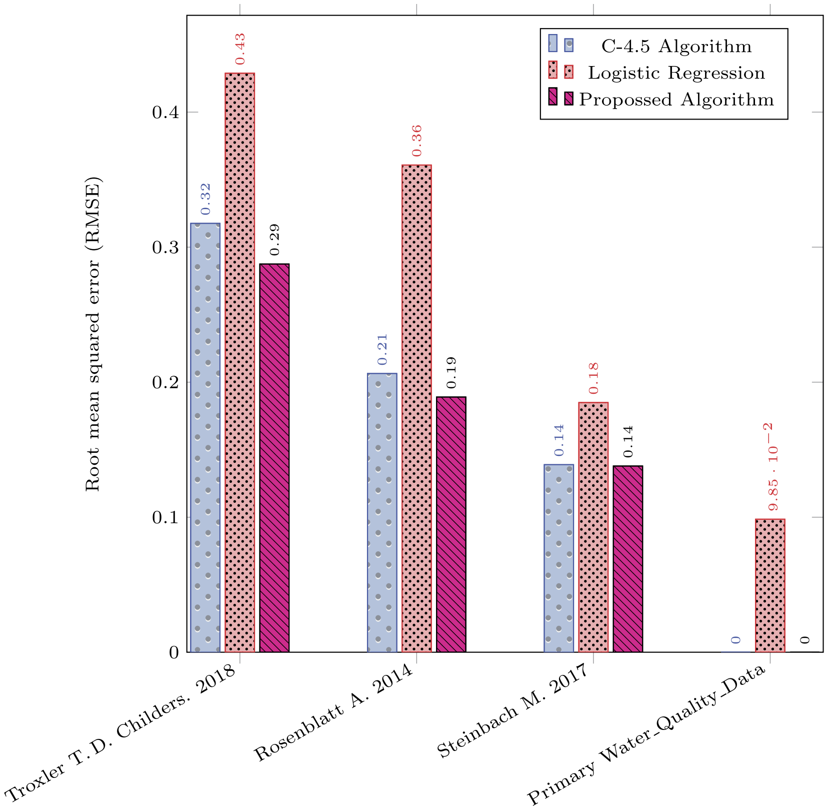


Fig. 9. Root mean squared error of C-4.5 algorithm, logistic regression & proposed algorithm for the dataset of Table 1.

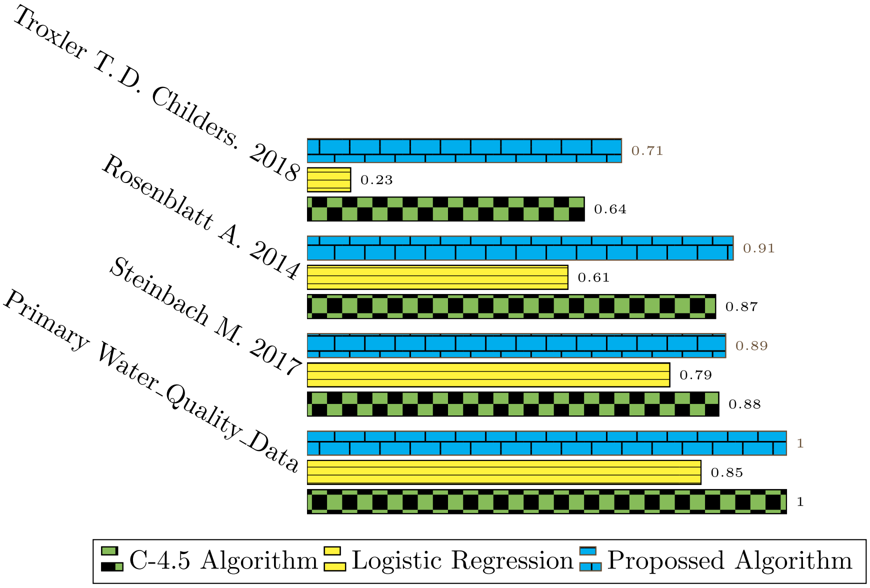


Fig. 10. Kappa statistic of C-4.5 algorithm, logistic regression & proposed algorithm for the dataset of Table 1.

duce higher precision if in the test dataset there are maximum instances of the same class.

$$K = \frac{P_0 - P_c}{1 - P_c} \tag{16}$$

Where P_0 is the total agreement probability i.e. Accuracy (Acc) as in equation (10)), and P_c is the probability that the agreement occurred by chance i.e. Random Accuracy (RA). The Random Accuracy is characterized as the aggregate of the products of reference probability and result probability for each class as in equation (17) in terms of $N_{\text{TP}}, N_{\text{TN}}, N_{\text{FP}}$ *and* N_{FN}.

$$\text{RA} = \frac{(N_{\text{TP}} + N_{\text{FP}})(N_{\text{TP}} + N_{\text{FN}})(N_{\text{TN}} + N_{\text{FP}})(N_{\text{TN}} + N_{\text{FN}})}{\left(N_{\text{TP}} + N_{\text{TN}} + N_{\text{FP}} + N_{\text{FN}}\right)^2} \tag{17}$$

If the number of class is more than two, we can represent Kappa Statistic as far as the confusion matrix's cell counts [24] is shown in equation (18).

$$K = \frac{R * \sum_{i=1}^{C} x_{ii} - \sum_{i=1}^{C} x_{i.}x_{.i}}{R^2 - \sum_{i=1}^{C} x_{i.}x_{.i}} \tag{18}$$

Where x_{ii} is the count of cases in the main diagonal, R is the number of instances, C is the number of classes, and $x_{i.}$ and $x_{.i}$ are the column and row total counts, respectively.

Comparison of the Kappa Statistic is shown in Fig. 10. The proposed algorithm gives a better Kappa Statistic value. Kappa Statistic is a normalised statistic. Its esteem never surpasses one, so a similar measurement can be utilised even when the number of occurrences go higher.

6. Conclusion

We have developed a water quality monitoring device consisting of existing sensors to collect the water quality data from three different sources. However, we have only two water quality parameters in our collected data. So, to validate the proposed method we have used three secondary water quality dataset having more quality parameters. We propose a classification method to classify the collected water quality data as well as to detect water contamination. We have also given a comparative study of the proposed method with two existing stander classification methods, i.e. C-4.5 and Logistic Regression. In both cases, we have experienced better classification performance by our reference indicator based decision tree algorithm. Thus, giving us a better water quality monitoring system.

Declarations

Author contribution statement

Swapan Shakhari: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Indrajit Banerjee: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by the Department of Science and Technology, Government of India (Project code: DST/TM/WTI/2K16/45(G)).

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2019.e01822.

References

[1] C.A. Varotsos, V.F. Krapivin, Pollution of Arctic waters has reached a critical point: an innovative approach to this problem, Water Air Soil Pollut. 229 (11) (2018) 343.

[2] F. Mkrtchyan, C. Varotsos, A new monitoring system for the surface marine anomalies, Water Air Soil Pollut. 229 (8) (2018) 273.

[3] C.A. Varotsos, V.F. Krapivin, A.A. Chukhlantsev, Microwave polarization characteristics of snow at 6.9 and 18.7 GHz: estimating the water content of the snow layers, J. Quant. Spectrosc. Radiat. Transf. 225 (2019) 219–226. URL http://www.sciencedirect.com/science/article/pii/S0022407318307726.

[4] Swapan Shakhari, Indrajit Banerjee, Water Quality (pH, TDS) Data – Collected from three different water sources namely “Packaged drinking water”, “Pipeline drinking water” and “Pond water” by a water quality monitoring device, Mendeley Data, v2, 2019.

[5] T. Troxler, D. Childers, Water Quality Data (Grab Samples) from the Taylor Slough, Everglades National Park (FCE), from May 2001 to Present, Environmental Data Initiative, 2018.

[6] A. Rosenblatt, Water Temperature, Salinity and other physical measurements taken at Shark River, Everglades National Park (FCE) from February 2010 to Present, Environmental Data Initiative, 2014. (Accessed 29 May 2019).

[7] M. Steinbach, Wisconsin Forest County Potawatomi Communit Water Resources Program lake chemistry 2003–2010, Environmental Data Initiative, 2017. (Accessed 29 May 2019).

[8] K.M.K. Kut, A. Sarswat, J. Bundschuh, D. Mohan, Water as key to the sustainable development goals of South Sudan–a water quality assessment of eastern Equatoria state, Groundwater Sustain. Dev. 8 (2019) 255–270.

[9] T. Mitrović, D. Antanasijević, S. Lazović, A. Perić-Grujić, M. Ristić, Virtual water quality monitoring at inactive monitoring sites using Monte Carlo optimized artificial neural networks: a case study of Danube river (Serbia), Sci. Total Environ. 654 (2019) 1000–1009.

[10] E. Fijani, R. Barzegar, R. Deo, E. Tziritis, S. Konstantinos, Design and implementation of a hybrid model based on two-layer decomposition method coupled with extreme learning machines to support real-time environmental monitoring of water quality parameters, Sci. Total Environ. 648 (2019) 839–853.

[11] R. Wan, F. Meng, E. Su, W. Fu, Q. Wang, Development of a classification scheme for evaluating water quality in marine environment receiving treated municipal effluent by an integrated biomarker approach in *Meretrix* meretrix, Ecol. Indic. 93 (2018) 697–703.

[12] H. Yousefi, S. Zahedi, M.H. Niksokhan, Modifying the analysis made by water quality index using multi-criteria decision making methods, J. Afr. Earth Sci. 138 (2018) 309–318.

[13] I.C. Nnorom, U. Ewuzie, S.O. Eze, Multivariate statistical approach and water quality assessment of natural springs and other drinking water sources in Southeastern Nigeria, Heliyon 5 (1) (2019) e01123.

[14] P. Kumar, Simulation of Gomti River (Lucknow City, India) future water quality under different mitigation strategies, Heliyon 4 (12) (2018) e01074.

[15] V. Roth, T. Lemann, G. Zeleke, A.T. Subhatu, T.K. Nigussie, H. Hurni, Effects of climate change on water resources in the upper Blue Nile Basin of Ethiopia, Heliyon 4 (9) (2018) e00771.

[16] A. Awotwi, M.A. Bediako, E. Harris, E.K. Forkuo, Water quality changes associated with cassava production: case study of white Volta basin, Heliyon 2 (8) (2016) e00149.

[17] J.R. Quinlan, C4. 5: Programs for Machine Learning, Elsevier, 2014.

[18] B. Krishnapuram, L. Carin, M.A.T. Figueiredo, A.J. Hartemink, Sparse multinomial logistic regression: fast algorithms and generalization bounds, IEEE Trans. Pattern Anal. Mach. Intell. 27 (6) (2005) 957–968.

[19] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30. URL http://dl.acm.org/citation.cfm?id=1248547.1248548.

[20] S. Garcia, F. Herrera, An extension on “statistical comparisons of classifiers over multiple data sets” for all pairwise comparisons, J. Mach. Learn. Res. 9 (Dec 2008) 2677–2694.

[21] C.A. Varotsos, C.G. Tzanis, N.V. Sarlis, On the progress of the 2015–2016 El Niño event, Atmos. Chem. Phys. 16 (4) (2016) 2007–2011.

[22] N. Tazin, S.A. Sabab, M.T. Chowdhury, Diagnosis of chronic kidney disease using effective classification and feature selection technique, in: 2016 International Conference on Medical Engineering, Health Informatics and Technology, MediTec, 2016, pp. 1–6.

[23] N. Kumar, S. Mitra, M. Bhattacharjee, L. Mandal, Comparison of different classification techniques using different datasets, in: Proceedings of International Ethical Hacking Conference 2018, Springer, 2019, pp. 261–272.

[24] A. Ben-David, What’s wrong with hit ratio? IEEE Intell. Syst. 21 (6) (2006) 68–70.

[25] A. Burgess, R. Boyd, J. Ziviani, M.D. Chatfield, R.S. Ware, L. Sakzewski, Stability of the Manual Ability Classification System in young children with cerebral palsy, Dev. Med. Child. Neurol. (2019).