Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

Enhanced Gradient Boosting Tree Classifier Using Optimization Technique for Water Quality Prediction

M. Durairaj[1] and T. Suresh[2]

[1] Assistant Professor, [2] Research Scholar
[1,2] School of Computer Science & Engineering, Bharathidasan University, Trichy, Tamilnadu, India
suresht24@gmail.com

ABSTRACT

Reservoirs and lakes are significant sources of water. Reservoirs are important water sources for all living things. They provide healthy drinking water as well as shelter for a wide range of marine organisms. Water from such resources may be used for a variety of purposes, including manufacturing, agriculture, drinking water supplies, recreation, and aesthetic value. Reservoirs are also useful for hydroelectric power production, flood control, and scenic beauty. Water stored in such tools can also be used during a drought. Unfortunately, these essential resources are contaminated, and water quality is affected by a number of factors. Anthropogenic practices, indiscriminate sewage disposal, human activities, and industrial waste all degrade water quality. Reservoir water quality control is important for the protection of marine resources. Water quality plays a role in regulating biotic diversity, biomass, energy, and succession rate. Furthermore, polluted water can cause certain waterborne illnesses and has an effect on child mortality. It is important to evaluate various aspects of water quality in order to reduce the effects of polluted water. It could be helpful to do this by forecasting water quality parameters a few steps ahead of time. The main aim of this paper is to come up with an effective classification method for predicting water quality using variable data. In this paper, the Differential Evolution (DE) optimization algorithm is used to improve the Gradient Boosting Tree classifier.

KEYWORDS: Water Quality, Classification, Prediction, Optimization technique, Differential Evolution, Gradient Boosting Tree.

1. INTRODUCTION

The most valuable resource for all of mankind is water. Lakes and reservoirs are examples of water bodies whose wise use may have a direct effect on humanity because all living organisms depend on them [1][2]. As a result, water bodies have been recognized as a domain for planning and management all over the world. Unfortunately, humans are losing

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

these important resources. Anthropogenic practices, environmental emissions from industry, sediments, and other factors all have a significant effect on water quality. The exploitation and deterioration of water sources has increased as a result of increased urbanization. Contaminated water supplies may have significant implications for both human and marine life. Aside from these, water quality is determined by a number of factors such as human activities, soils, geology, and so on. Furthermore, indiscriminate waste disposal and human activities degrade water quality. Contaminated water can also cause waterborne illnesses and has an effect on infant mortality. It is important to evaluate various aspects of water quality in order to reduce the effects of polluted water. It could be helpful to do this by forecasting water quality parameters a few steps ahead of time. As a result, reservoir water quality control is important for utilization and protection of aquatic resources [3][4]. Water quality plays a role in regulating biotic diversity, biomass, energy, and succession rate.

For decades, water contamination has been a major problem in many developing countries. Water shortage is another big problem that can be caused by polluted water [5]. Water quality can be influenced by a variety of factors. Thermal contamination, acidification, and salinization are only a few examples. Furthermore, surface water temperatures are influenced by urbanization and industrial effluents, which influences dissolved oxygen levels. The metabolism of the species living in the reservoir relies on dissolved oxygen. The pH of water may be affected by industrial effluents or human activities in the vicinity of the water source. In comparison, Salinization decreases the amount of safe drinking water available and threatens marine life. Turbidity is a property of water that determines its purity. A higher degree of transparency in water bodies suggests that the water is of good quality. During the monsoon, when the water is murkier and suspended solids are present, reservoirs have a high turbidity. Water quality is determined by temperature. It has an effect on the speed of chemical reactions and aquatic metabolism. Since aquatic metabolism has a poor temperature tolerance, even a slight change in the temperature of the surface water will harm the metabolism of the organisms living in the water basin. Nutrients are important for life to survive. Water quality can have an effect on nutrients as well as aquatic metabolism [6]. The knowledge obtained from water quality monitoring can then be used to make quality forecasts a few steps ahead of time in order to improve water management [7].

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

2. RELATED WORKS

Baek, Sang-Soo, JongcheolPyo, and Jong Ahn Chun [8] Combining CNN and LSTM networks, researchers developed a Convolutional Neural Network (CNN)-Long Short-Term Memory (LSTM) coupled with a deep learning approach to simulate water quality, including total nitrogen, total phosphorous, and total organic carbon. The Water Resources Management Information System (WAMIS) and Real-TimeWater Quality Information, respectively, were used to collect water level and water quality data in the Nakdong river basin.

Yan, Jianzhuo, et al [9] The deep belief network (DBN) model was used to propose a framework for forecasting water quality. The PSO algorithm is first used to optimize the network parameters of the deep belief network, which is used to extract feature vectors of water quality time series data at multiple scales. The PSO-DBN-LSSVR water quality prediction model is then combined with the least squares support vector regression (LSSVR) machine, which is used as the top prediction layer of the model.

Fu, Zhao [10] Several methods for improving the performance of Adaptive Neuro Fuzzy Inference System (ANFIS)-based water quality prediction models are proposed in this dissertation. In the training and testing datasets, stratified sampling is used to account for various types of data distribution. The wavelet was used to filter out the noise in the dataset. After applying stratified sampling and wavelet denoising techniques, a deep prediction performance comparison between Multivariate Linear Regression (MLR), Artificial Neural Network (ANN), and ANFIS model is provided.

Lu, Hongfang, and Xin Ma [11] To obtain more reliable short-term water quality prediction performance, two novel hybrid decision tree-based machine learning models were proposed. Extreme gradient boosting (XGBoost) and random forest (RF) are the basic models of the two hybrid models, which implement an advanced data denoising technique - full ensemble empirical mode decomposition with adaptive noise - respectively (CEEMDAN).

Asadollah, Seyed Babak Haji Seyed, et al [12] Extra Tree Regression (ETR) is a new ensemble machine learning model for predicting monthly WQI values at the Lam Tsuen River in Hong Kong. The performance of the ETR model is compared to that of conventional standalone models such as Support Vector Regression (SVR) and Decision Tree Regression (DTR) (DTR). The prediction models are developed using monthly input water quality data such as Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Dissolved

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

Oxygen (DO), Electrical Conductivity (EC), Nitrate-Nitrogen (NO3-N), Nitrite-Nitrogen (NO2-N), Phosphate (PO4 3), Hydrogen Potential (pH), Temperature (T), and Turbidity (TUR).

Mangai, J. Alamelu, and Bharat B. Gulyani [13] presented a data-driven model for predicting BOD in a lower-dimensional space using dimensionality reduction techniques to eliminate irrelevant properties from high-dimensional data. The complete dataset with 11 parameters was used to train machine learning algorithms such as decision stump, SVM, MLP, linear regression (LR), and instance-based learner (IBK).

Liu, Cong, Hongji Li, and Qinkun Zhang [14] proposed a sensor-based wireless dynamic water quality monitoring and prediction system. First, this paper establishes a sewage control model and real-time dynamic monitoring of total nitrogen, total phosphorus, ammonia nitrogen, and other measures of the basin's water quality using wireless sensor technology and the ZigBee protocol. Second, a support vector algorithm is used to create a water quality prediction model based on wireless monitoring in order to make a fair prediction of the watershed's water quality.

Wu, Di, Hao Wang, and Razak Seidu [15] proposed a smart data analysis scheme for monitoring and predicting water quality, taking into account all of the water quality level indicators The authors obtained raw water data directly from water sources rather than using data from water treatment. To predict water quality, the authors developed two models: (1) adaptive learning rate BP neural network (ALBP) and (2) 2-step isolation and random forest (2sIRF). These models were tested in the real-world urban water supply systems of Oslo and Bergen, Norway.

3. ENHANCED GRADIENT BOOSTING TREE CLASSIFIER

3.1 Differential Evolution

A well-known evolution-based optimization technique is differential evolution [16]. Genetic Algorithm (GA) mutation and crossover operators have been elegantly merged. The most important distinction between GA and differential evolution is that GA needs less control parameters. As a consequence, it has a higher convergence rate than GA. The method of parameter selection can be supported by differential evolution. Differential evolution, in general, consists of four main steps to find suitable solutions (i.e. initializing population, mutation and recombination, selection, and stopping criteria) (see Figure 1). These measures are briefly discussed in the following section.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

(i)Initializing population: In this step, a set of random solutions is created, with each parameter falling within the lower and upper bounds. These random values are generated using the normal distribution $\left(ND(\mu, \sigma^2)\right)$.

(ii)Mutation and recombination: The way differential evolution produces solutions during its life-cycle is its primary advantage. The discrepancy between any two differential evolution solutions is combined with the third solution to form a new solution. As a consequence, it can elegantly substitute GA's mutation and crossover operators. This shows that differential evolution can converge faster than GA.

(iii) Selection: The fitness value of solutions is then measured using a predetermined objective function. If the fitness of the new solution exceeds that of the best known solution so far, the best known solution will be replaced with the new solution, and the old solution will be retained for future generations.

(iv) Stopping criteria: Maximum fitness function is difficult to achieve. To complete the differential evolution process, one may use the number of iterations, the number of function evaluations, or the acceptance error (AE).

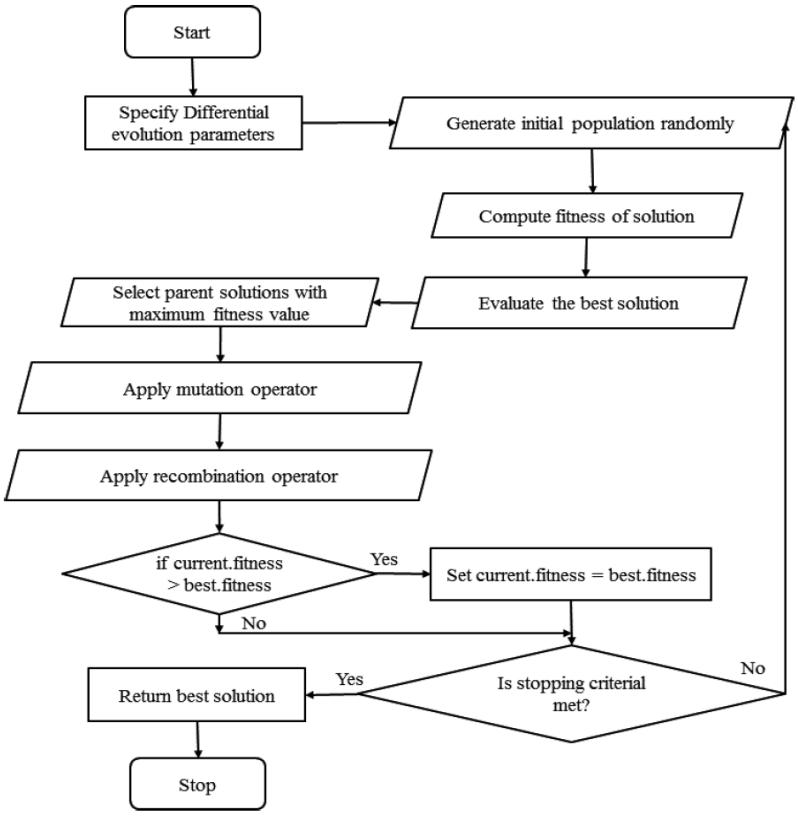

Figure 1: Flowchart of Differential Evolution

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

3.2 Gradient Boosting Tree

Gradient boosting [17] is a strong ensemble machine learning technique for classification and regression problems that combines a series of weak prediction models, usually decision trees, to generate a classification or regression model. The GBDT algorithm uses gradient boosting to expand and improve the classification and regression tree model. Unlike the random forest algorithm, the GBDT learning technique fits new models in a sequential manner to provide a more accurate estimate of the response variables. The GBDT algorithm creates decision trees iteratively. A decision tree is learned from the residuals of the previous tree in each iteration. Finally, the output is generated by combining the classified results of all trees.

3.3 Proposed Enhanced Gradient Boosting Tree Classifier with Differential Evolution Optimization

In this research work, Gradient Boosting Tree classifier can be enhanced by utilizing Differential Evolution (DE) optimization algorithm. A lot of researchers have applied DE for the parameter optimization. DE uses Unlike GA and PSO, DE uses difference of randomly sampled pairs of object vectors to guide the mutation process, and it is a parallel direct search method introduced by Storn and Price.DE can be categorized into a class of floating-point encoded, evolutionary optimization algorithms, and it has been widely applied by researchers in the last decades.

Step by Step Procedure for Proposed Enhanced Gradient Boosting Tree Classifier

Input: Learning dataset, Number of iterations, Loss function, and learning rate.

Output: The predict model

Step 1: Initialization of the GBDT parameters through DE parameters.

Step 2: Calculate the Negative Gradient of the loss Function.

Step 3: Build a new decision tree. Linear search is utilized to minimize the loss function and to estimate the value of the each leaf node region.

Step 4: Updation of the regression tree. Then the updated regression value is send to the fitness evaluation of each population is evaluated of DE.

Step 5: Applying mutation operator in DE self-referential population recombination scheme.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

Step 6: Applying Uniform crossover operator.

Step 7: Applying fitness evaluation and selection is applied to improve the solutions by selection process.

Step 8: After initialization, the process of mutation, crossover and selection is repeated until a predetermined generation number is obtained or a termination criterion is satisfied.

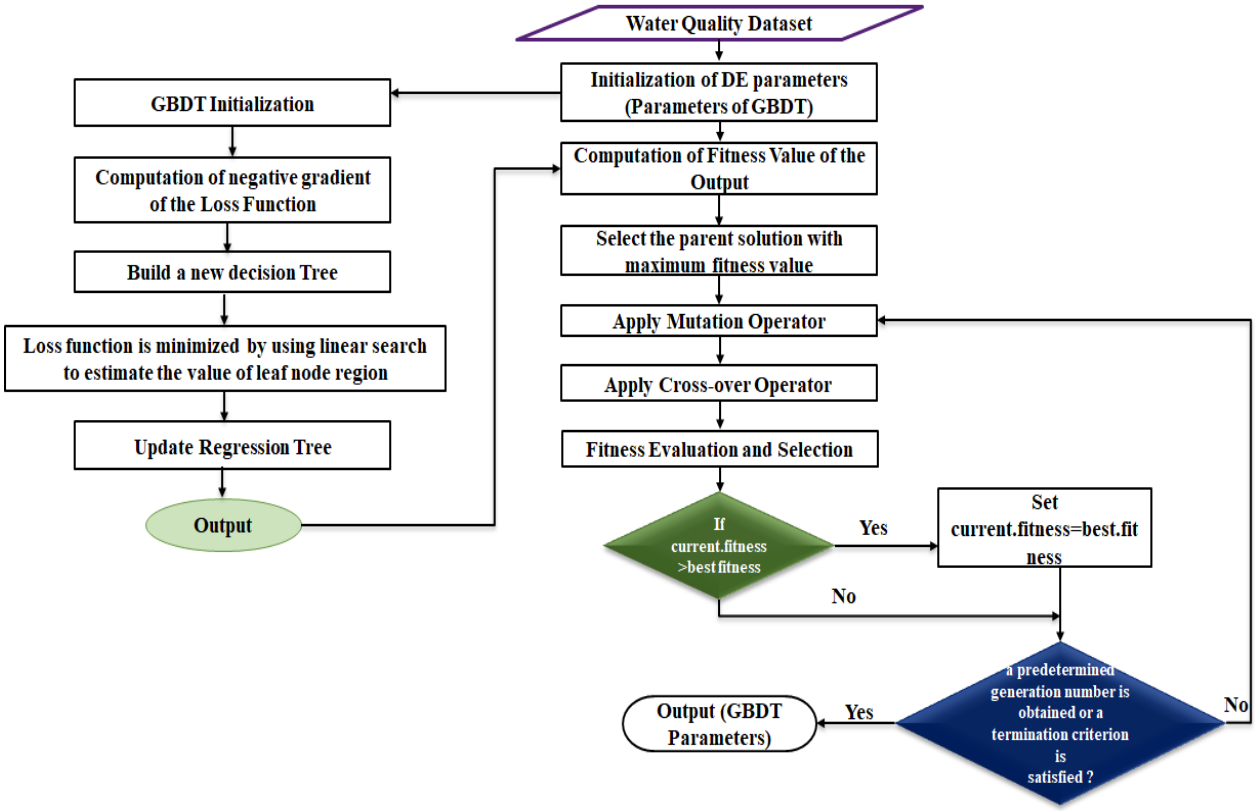

Figure 2: Proposed Enhanced Gradient Boosting Tree Classifier Flowchart

4. RESULT AND DISCUSSION

4.1 Performance Metrics

The following table 1 depicts the performance metrics considered in this research work. To evaluate the performance of the proposed classifier and it is compared with the existing classifiers like Random Forest, Gradient Boosting Tree (GBT), and K-Nearest Neighbor. The performance of the proposed classifier is evaluated using Random Forest (Feature of Importance) and Gradient Boosting Tree (Feature of Importance) as the pre-processing methods.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

Table 1: Performance Metrics used in this research work

Metrics	Equation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision	$\frac{TP}{TP+FP}$
False Positive Rate	1-Specificity
Miss Rate	1-Sensitivity
False Discovery Rate	1-Precision

4.2 Description of the Dataset

Table 2 depicts the description of the dataset used in this research work.

Table 2: Description of the dataset

Feature Number	Feature Name	Description
1	Station Code	The code used to represent the water station code
2	Locations	The exact location of the water source used for the quality prediction
3	State	The States in India where the water source located
4	D.O. (mg/l)	Dissolved oxygen is usually reported in milligrams per liter (mg/L) or as a percent of air saturation
5	PH	pH is a measure of how acidic/basic water is
6	CONDUCTIVITY (µmhos/cm)	Conductivity of water is measured in micromhos per centimeter (µmhos/cm) or microsiemens per centimeter (µs/cm)
7	B.O.D. (mg/l)	Biochemical Oxygen Demand is a measure in

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

		Milligrams per liter is the amount of oxygen in a liter of water
8	NITRATENAN N+ NITRITENANN (mg/l)	It's a nitrate concentration as milligrams per liter (mg/L) in water
9	FECAL COLIFORM (MPN/100ml)	Used in water microbiology to denote coliform organisms. It is measured in Most Probable Number (MPN) per 100 milli litre
10	TOTAL COLIFORM (MPN/100ml)Mean	Total Coliform, Fecal coliform E. coli. bacterial contamination of drinking water
11	Year	The year when the water sample taken for testing
12	Class	Two categories (Yes: Used for drinking (Quality approved), No: Not suggested for drinking (No Quality approved))

4.3 Number of Features obtained

Table 3 gives the number of features obtained by the Pre-processing techniques like RF (FOI) and GBT(FOI).

Table 3: Number of features obtained

Pre-Processing Methods	Number of Features obtained
Original Dataset without pre-processing methods	11
RF (FOI)	10
GBT (FOI)	9

4.4 Performance analysis of the Proposed Enhanced Gradient Boosting Tree classifier

Table 4 depicts the Classification Accuracy (in %) of proposed E-GBDT, RF, GBT and KNN classifiers for given Original dataset, RF(FOI), BGT(FOI). Table 5 gives the Sensitivity (in %) of proposed E-GBDT, RF, GBT and KNN classifiers for given Original dataset, RF(FOI), and GBT(FOI). Table 6 presents the Specificity (in %) of proposed E-GBDT, RF, GBT and KNN classifiers for given Original dataset, RF(FOI), and GBT(FOI).

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

Table 7 depicts the Precision (in %) of proposed E-GBDT, RF, GBT and KNN classifiers for given Original dataset, RF(FOI), and GBT(FOI). Table 8 presents the False Positive Rate (in %) of proposed E-GBDT, RF, GBT and KNN classifiers for given Original dataset, RF(FOI), and GBT(FOI). Table 9 gives the Miss Rate (in %) of proposed E-GBDT, RF, GBT and KNN classifiers for given Original dataset, RF(FOI), and GBT(FOI). Table 10 depicts the False Discovery Rate (in %) of proposed E-GBDT, RF, GBT and KNN classifiers for given Original dataset, RF(FOI), and GBT(FOI).

From the table 4, table 5, table 6, and table 7, it is clear that the classification accuracy, sensitivity, specificity, and precision is increased with pre-processing method as GBT(FOI) using Proposed E-GBT classifier than the other classification techniques.

From the table 8, table 9 and table 10, it is clear that the FPR, Miss Rate and FDR are reduced using proposed E-GBT with GBT (FOI) as the pre-processing method than the other methods and classification techniques.

Table 4: Classification Accuracy (in %) of the Proposed E-GBDT, RF, GBT and KNN classifier for given Original dataset, RF(FOI), BGT(FOI)

Pre-Processing Methods	Classification Techniques			
	Proposed E-GBT	**RF**	**GBT**	**KNN**
Original Dataset	57.26	32.22	35.44	27.33
RF (FOI)	63.54	46.22	51.00	41.66
GBT(FOI)	80.67	65.22	72.00	58.88

Table 5: Sensitivity (in %)of the Proposed E-GBDT, RF, GBT and KNN classifier for given Original dataset, RF(FOI), BGT(FOI)

Pre-Processing Methods	Classification Techniques			
	Proposed E-GBT	**RF**	**GBT**	**KNN**
Original Dataset	56.87	33.80	36.55	28.20
RF (FOI)	62.26	46.51	50.88	42.33
GBT(FOI)	64.58	62.52	67.24	57.09

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

Table 6: Specificity (in %)of the Proposed E-GBDT, RF, GBT and KNN classifier for given Original dataset, RF(FOI), BGT(FOI)

Pre-Processing Methods	Classification Techniques			
	Proposed E-GBT	RF	GBT	KNN
Original Dataset	42.74	24.30	24.60	21.42
RF (FOI)	68.36	36.91	40.77	32.87
GBT(FOI)	77.32	54.08	61.50	47.77

Table 7: Precision (in %)of the Proposed E-GBDT, RF, GBT and KNN classifier for given Original dataset, RF(FOI), BGT(FOI)

Pre-Processing Methods	Classification Techniques			
	Proposed E-GBT	RF	GBT	KNN
Original Dataset	**43.27**	**30.36**	**32.36**	**24**
RF (FOI)	69.31	41.27	46.90	37.63
GBT(FOI)	73.68	62.18	70.18	58.54

Table 8: False Positive Rate (in %)of the Proposed E-GBDT, RF, GBT and KNN classifier for given Original dataset, RF(FOI), BGT(FOI)

Pre-Processing Methods	Classification Techniques			
	Proposed E-GBT	RF	GBT	KNN
Original Dataset	57.26	75.7	75.4	78.58
RF (FOI)	31.64	63.09	59.23	67.13
GBT(FOI)	22.68	45.92	38.5	52.23

Table 9: Miss Rate (in %)of the Proposed E-GBDT, RF, GBT and KNN classifier for given Original dataset, RF(FOI), BGT(FOI)

Pre-Processing Methods	Classification Techniques			
	Proposed E-GBT	RF	GBT	KNN
Original Dataset	43.13	66.2	63.45	71.8
RF (FOI)	37.74	53.49	49.12	57.67

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

GBT(FOI)	35.42	37.48	32.76	42.91

Table 10: False Discovery Rate (in %)of the Proposed E-GBDT, RF, GBT and KNN classifier for given Original dataset, RF(FOI), BGT(FOI)

Pre-Processing Methods	Classification Techniques			
	Proposed E-GBT	**RF**	**GBT**	**KNN**
Original Dataset	56.73	69.64	67.64	76
RF (FOI)	30.69	58.27	53.04	62.37
GBT(FOI)	26.32	37.82	29.82	41.46

5. CONCLUSION

Water is the most valuable of all commodities, necessary for the survival of all types of life; however, it is continually polluted by life itself. Water is one of the most communicable and far-reaching mediums. As a result of rapid industrialization, water quality has degraded at an alarming pace. To improve the classification and prediction accuracy of water quality, an enhanced Gradient Boosting Tree classifier was designed with Differential Evolution optimization in this report. It can be seen from the results that the proposed E-GBDT classifier outperforms existing classifiers such as KNN, GBT, and RF. It is apparent that the proposed E-GBDT classifier outperforms the other current classifiers.

REFERENCES

[1] Ahmed, Ali Najah, et al. "Machine learning methods for better water quality prediction." *Journal of Hydrology* 578 (2019): 124084.

[2] Muharemi, Fitore, DoinaLogofătu, and Florin Leon. "Machine learning approaches for anomaly detection of water quality on a real-world data set." *Journal of Information and Telecommunication* 3.3 (2019): 294-307.

[3] Chen, Yingyi, et al. "A review of the artificial neural network models for water quality prediction." *Applied Sciences* 10.17 (2020): 5776.

[4] Arefinia, Ali, et al. "Reservoir water quality simulation with data mining models." *Environmental Monitoring and Assessment* 192.7 (2020): 1-13.

[5] Aldhyani, Theyazn HH, et al. "Water Quality Prediction Using Artificial Intelligence Algorithms." *Applied Bionics and Biomechanics* 2020 (2020).

[6] Najafzadeh, M., A. Ghaemi, and S. Emamgholizadeh. "Prediction of water quality parameters using evolutionary computing-based formulations." *International Journal of Environmental Science and Technology* 16.10 (2019): 6377-6396.

[7] Ahmed, Ali Najah, et al. "Machine learning methods for better water quality prediction." *Journal of Hydrology* 578 (2019): 124084.

[8] Baek, Sang-Soo, JongcheolPyo, and Jong Ahn Chun. "Prediction of Water Level and Water Quality Using a CNN-LSTM Combined Deep Learning Approach." *Water* 12.12 (2020): 3399.

[9] Yan, Jianzhuo, et al. "A Prediction Model Based on Deep Belief Network and Least Squares SVR Applied to Cross-Section Water Quality." *Water* 12.7 (2020): 1929.

[10] Fu, Zhao. *Water Quality Prediction Based on Machine Learning Techniques*. Diss. University of Nevada, Las Vegas, 2020.

[11] Lu, Hongfang, and Xin Ma. "Hybrid decision tree-based machine learning models for short-term water quality prediction." *Chemosphere* 249 (2020): 126169.

[12] Asadollah, Seyed Babak Haji Seyed, et al. "River water quality index prediction and uncertainty analysis: A comparative study of machine learning models." *Journal of Environmental Chemical Engineering* (2020): 104599.

[13] Mangai, J. Alamelu, and Bharat B. Gulyani. "Dimensionality Reduction for Water Quality Prediction from a Data Mining Perspective." *International conference on Modelling, Simulation and Intelligent Computing*. Springer, Singapore, 2020.

[14] Liu, Cong, Hongji Li, and Qinkun Zhang. "Research on Sewage Monitoring and Water Quality Prediction Based on Wireless Sensors and Support Vector Machines." *Wireless Communications and Mobile Computing* 2020 (2020).

[15] Wu, Di, Hao Wang, and Razak Seidu. "Smart data driven quality prediction for urban water source management." *Future Generation Computer Systems* 107 (2020): 418-432.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3860 - 3873
Received 20 January 2021; Accepted 08 February 2021.

[16] Wang, Zi-Jia, et al. "Automatic niching differential evolution with contour prediction approach for multimodal optimization problems." *IEEE Transactions on Evolutionary Computation* 24.1 (2019): 114-128.

[17] Lu, Hongfang, et al. "Short-term prediction of building energy consumption employing an improved extreme gradient boosting model: a case study of an intake tower." *Energy* 203 (2020): 117756.