Recognizing Safe Drinking Water and Predicting Water Quality Index using Machine Learning Framework

Mohamed Torky[1], Ali Bakhiet[2], Mohamed Bakrey[3], Ahmed Adel Ismail[4], Ahmed I. B. EL Seddawy[5]
Faculty of Artificial Intelligence, Egyptian Russian University (ERU), Badr City, Egypt[1]
Higher Institute of Computer Science and Information Systems, Culture & Science City, Giza, Egypt[2, 3]
The Higher Institute of Computer and Information Systems, Abo Qir Alexandria 21913, Egypt[4]
Arab Academy for Science and Technology and Maritime Transport, Cairo, Egypt[5]

Abstract*—Water quality monitoring, analysis, and prediction have emerged as important challenges in several uses of water in our life. Recent water quality problems have raised the need for artificial intelligence (AI) models for analyzing water quality, classifying water samples, and predicting water quality index (WQI). In this paper, a machine-learning framework has been proposed for classify drinking water samples (safe/unsafe) and predicting water quality index. The classification tier of the proposed framework consists of nine machine-learning models, which have been applied, tested, validated, and compared for classifying drinking water samples into two classes (safe/unsafe) based on a benchmark dataset. The regression tier consists of six regression models that have been applied to the same dataset for predicting WQI. The experimental results clarified good classification results for the nine models with average accuracy, of 94.7%. However, the obtained results showed the superiority of Random Forest (RF), and Light Gradient Boosting Machine (Light GBM) models in recognizing safe drinking water samples regarding training and testing accuracy compared to the other models in the proposed framework. Moreover, the regression analysis results proved the superiority of LGBM regression, and Extra Trees Regression models in predicting WQI according to training, testing accuracy, 0.99%, and 0.95%, respectively. Moreover, the mean absolute error (MAE) results proved that the same models achieved less error rate, 10% than other applied regression models. These findings have significant implications for the understanding of how novel deep learning models can be developed for predicting water quality, which is suitable for other environmental and industrial purposes.

Keywords*—*Water quality; artificial intelligence; machine learning; deep learning; classification analysis; and regression analysis

I. Introduction

In the new green economy, monitoring and evaluating water quality is a central issue for the life of all organisms. Using the classical monitoring ways that depend on chemical monitoring is not enough to evaluate the consequences of some influences and stresses, as predicting the interactive effects of different chemical variables on water microorganisms is very difficult [1]. Rapid industrial development has deteriorated water quality at an alarming rate. In addition, the infrastructure, with the absence of public awareness, and the low quality of hygiene, greatly affects the quality of drinking water [2]. Polluted drinking water is very serious and can adversely affect organisms' health, as well as many environmental, and infrastructural impacts. According to a United Nations (UN) report, roughly, more than 1.5 million people die every year due to water-polluted diseases. In third-world countries, it has been declared that 80% of health issues are due to polluted water. Moreover, 2.5 billion illnesses and five million deaths are reported annually [3], and these are truly terrifying numbers.

Due to the lack of robust water monitoring techniques, many countries are unable to enhance their water systems and there are shortcomings to produce effective water recovery systems. These shortcomings may lead to a greater level of uncertainty when developing water resource management policies [4].

Recently, there has been a marked increase in the development of rapidly developing biological monitoring and biological assessment tools for water resources that are reliable enough to manage many degraded water bodies in the USA, Europe, South Africa, and Australia [5]. However, with the huge increase in data generated by monitoring devices and the futility of manual coding, the shortcomings began to appear in those systems due to the lack of an effective mechanism for processing that huge data. However, with the growth of artificial intelligence based on machine learning and deep learning techniques, it can introduce a perfect solution to that problem, such as artificial intelligence is characterized by many predictions, clustering, and classification techniques to produce effective solutions to water quality problems [6]. Research of the past decades has focused largely on analyzing the water quality of rivers based on artificial intelligence (AI) techniques [7]. Using AI models, water quality forecasting, classification, and risk assessment can be achieved easily. Moreover, advanced early warning systems and effective management policies can be designed to add more control and monitoring services to rivers and water bodies [8, 9].

In this paper, a proposed machine learning framework has been introduced for analyzing water quality. It consists of two subsystems; the first subsystem is responsible for classifying water quality based on nine AI models that have been applied, tested, and compared to classify various samples of drinking water as safe to drink or unsafe to drink. The applied nine AI models are: Extreme Gradient Boosting (XGBoost) [10], Light Gradient Boosting Machine (Light GBM) [11], Decision Tree (DT) [12], Extra Tree (ET) [13], Multi-layer Perceptron (MLP) [14], Gradient Boosting (GB) [15], Support Vector Machine (SVM) [16], Artificial Neural Network (ANN) classification [17], and Random Forest (RF) Classifier [18]. The second subsystem is responsible for predicting water quality index (WQI) based on six regression models, LGBM regression, XGB regression, ExtraTrees regression, DT Regression, RF regression, and linear regression. These models have been applied to a dataset called Water quality, which was downloaded from [19]. The experimental results proved the superiority of the LightGBM model compared with the other eight AI models with an accuracy of 97% in classifying water samples to recognize the safe drinking water samples. Moreover, the predictive analysis of the used regression models clarified outperforms of LGBM regression, and Extra Trees Regression models in predicting water quality index according to training accuracy, testing accuracy, and mean absolute error (MAE) compared to the other four regression models.

The rest of this article is designed as follows: Section II reviews the related work. Section III explains the proposed machine-learning framework for analyzing water quality. Section IV presents and discusses the implementation results. Section V presents the conclusion of this work.

II. Literature Review

A growing body of literature has investigated the efficiency of using machine and deep learning models for monitoring, analyzing, and predicting water quality index. The literature introduced some reviews that discuss various AI models for solving water quality prediction problems [9,20,21]. There are several large cross-sectional studies, which introduces multiple machine and deep learning to predict water quality index.

Ali Najah et al. [22] applied four machine learning models, an enhanced Wavelet De-noising Techniques (WDT)-based Neuro-Fuzzy Inference System (WDT-ANFIS), Adaptive Radial Basis Function Neural Networks (RBF-ANN), Neuro-Fuzzy Inference System (ANFIS), Multi-Layer Perceptron Neural Networks (MLP-ANN), and to predict water quality parameters (i.e. pH, ammonia nitrogen (AN), and suspended solids (SS)) of Johor River in Malaysia. The experimental results clarified outperform of the WDT-ANFIS model in prediction accuracy for all the water quality parameters compared to the other three used models.

Amir Hamzeh et al. [23] used the support vector machine (SVM) algorithm, Artificial Neural Network (ANN), and group method of data handling (GMDH) models for analyzing the water quality prediction of Tireh River in Iran. Different types of the kernel and transfer functions were validated and tested, and the practical results clarified that both ANN and SVM are better models than GMDH in predicting the water quality of Tireh River.

Umair Ahmed et al [24] introduced supervised learning models for evaluating WQI prediction based on four features of water elements, namely, turbidity, temperature, pH, and total dissolved solids. The proposed models achieved acceptable accuracy and fewer error rates using a minimal number of features in predicting the WQI in real-time.

Abubakr Saeed et al. [25] proposed an efficient machine learning algorithm based on the SVM model to forecast the WQI of Langat River Basin based on the investigation of six variables (Dissolved Oxygen (DO), pH, Chemical Oxygen Demand (COD), Suspended Solids (SS), Ammonia Nitrogen (AN), and Biochemical Oxygen Demand (BOD)) of dual reservoirs that are located in the catchment. The experimental results showed that this model could accurately predict WQI value with small mean absolute error.

Mourad Azrour et al. [26] investigated the efficiency of machine learning algorithms for evaluating WQI prediction value based on four water features: pH, temperature, turbidity, and coliforms. The experimental results have proven the efficiency of used regression algorithms in predicting WQI. Moreover, the artificial neural network proved that it is the most highly efficient model in classifying water quality compared to other models in the literature.

They H et al. [27] utilized advanced AI models to evaluate WQI prediction value and classifying water goodness. The authors applied nonlinear autoregressive neural networks (NARNET) and long short-term memory (LSTM) as deep learning algorithms for predicting WQI. Moreover, three learning techniques, namely, K-nearest neighbor (K-NN), Naive Bayes, and SVM have been applied for the water quality classification task. The Prediction results showed that the NARNET algorithm performed slightly better than the LSTM for predicting WQI values. On the other hand, the SVM model has achieved the greatest accuracy (97.01%) for water goodness classification compared to the other classification models.

Siti Nur Mahfuzah et al. [28] investigated the efficiency of two machine learning algorithms, the Random Forest algorithm and the Random Tree algorithm for Classifying River Water Quality. The practical results have proven that Random Forest gives a higher classification accuracy compared to the Random Tree algorithm.

Junhao Wu et al. [29] proposed a hybrid model based on discrete wavelet transform (DWT), an ANN model, and LSTM model to predict the water goodness of the Jinjiang River. The prediction results clarified the efficiency of the proposed hybrid model in predicting water quality index compared to other models such as the ARIMA model, the LSTM model, nonlinear autoregression (NAR) model, the ANN-LSTM model, multi-layer perceptron model, and the CNN-LSTM model.

NguyenHien Than et al. [30] investigated water quality monitoring for the Dong Nai River at different times based on a novel architecture of the neural network model FFNN, and LSTM-MA hybrid model at different time series. The validation results proved that The LSTM-MA model provided more reliable prediction and achieved faster training time than the NAR, NAR-MA, ARIMA, and LSTM models. Moreover, the proposed hybrid model produced classification results for water quality in close agreement with the actual monitoring data.

Other hybrid machines and deep learning models have been developed for investigating water quality index, for example, one-dimensional residual CNN (1-DRCNN) and bi-directional gated recurrent units (BiGRU) have been utilized for predicting Water Quality in the Luan River [31]. Moreover, a hybrid deep learning model based on the CNN and LSTM model has been applied, tested, and compared for predicting water goodness based on real-time monitoring of water quality variables [32].

III. Water Quality Analysis Framework

Automatic analyzing drinking water quality from a given dataset, a framework consisting of two phases is proposed. The first phase is responsible for classifying water samples from a given dataset into two classes, safe or unsafe for drinking based on nine classification algorithms, whereas, the second phase is responsible for predicting the water quality index (WQI) based on six regression algorithms. In the following, the two phases are discussed in more detail:

A. *Phase 1: Water Samples Classifications*

To classify water samples to recognize safe drinking water samples, nine-machine learning techniques have been used, tested, and compared. Fig. 1 depicts how these models can be used for classifying water samples from a given dataset. The classification phase starts by doing a preprocessing step for cleaning, splitting, and resampling the used dataset. In the second step, the given dataset is divided into training (70%) and testing (30%) data parts. The third step focuses on extracting water features that may impact water quality through a feature selection step. The final step, the classification step sequentially calls nine classification algorithms (i.e. learning model) one after one for performing the classification task. The used classification models can be briefly described as follows:

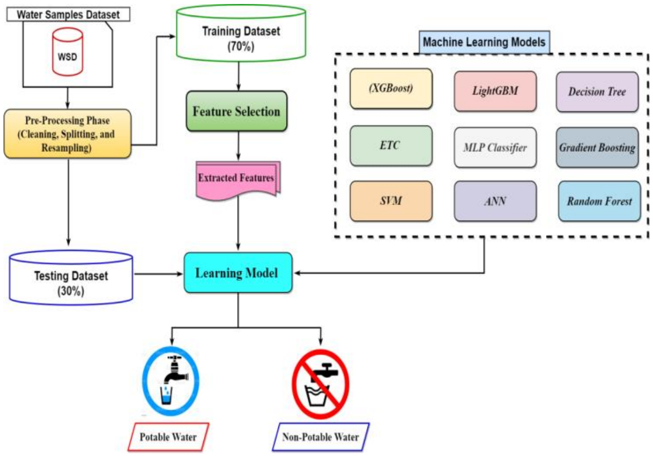


Fig. 1. Water quality classification model.

1) Extreme Gradient Boosting (XGBoost): It is depending on supervised machine learning, decision trees, ensemble learning, and gradient boosting. It is one of the most powerful techniques for building stochastic models for regression, classification, and ranking problems [33]. It provides a parallel tree boosting approach to fix errors made by prior boosted tree models [34].

2) Light Gradient Boosting Machine (Light GBM): It has been developed by Microsoft, which is a popular algorithm used for ranking and classification problems. Its structure is also based on decision tree models. LightGBM is being distinguished by training speed and accurate prediction results. This is because of adding an automatic feature selection procedure as well as focusing on boosting instances with greater gradients [35].

3) Decision Tree (DT): It is a common supervised learning algorithm used for regression and classification problems [12]. The idea is to use learning decision rules deduced from the data features to perform classification or prediction tasks. What makes DT an effective classification model is: 1) the DT model can be prepared with little data. 2) Training a DT model is logarithmic in the number of data points. 3) A DT model can be validated by statistical tests. 4) Its performance doesn't affect any violation in predefined assumptions with the original model from which the data were created. 5) DT models can be visualized easily and can be understood without mysterious [36].

4) Extra Tree Classifier (ETC): It is a class of ensemble learning approaches. The classification results are collected from a forest of several de-correlated DT models [37]. It differs from Random Forest Classifier in DT constructions way, where DT models are constructed in a "forest". The forest construction and creation of multiple de-correlated DT models of this classifier are based on extracting a random sample of features that leads to the best classification results based on some mathematical conditions.

5) Multi-layer Perceptron Classifier (MLP Classifier): It is a class of feed-forward neural network models [38]. There may be multiple nonlinear hidden layers between the input and the output layers for mapping input data to output data. This classifier is based on the functionality of the sigmoid activation function for doing the classification task.

6) Gradient Boosting Classifier (GBC): It is a common boosting classifier algorithm [39]. The functionality of gradient boosting works based on training N Trees based on the repeated fixing errors resulting from the predecessors of predictors to form the ensemble of data. The training step of the GBC model is done by training the predictors with the error labels produced by the predecessor of those predictors. The prediction results of each tree model are based on "a shrinking routine".

7) Support Vector Machine (SVM) Classifier: it is a supervised learning model used for both regression and classification problems [40]. The main goal of the SVM model is to identify a hyperplane in an N-dimensional space for classifying data items. The kernel of SVM is a procedure that depends on low-dimensional input space and converts it into higher-dimensional space. Therefore, SVM is suitable for non-linear classification problems. SVM has some advantages that make it an efficient classifier such as memory efficiency, effectiveness in high dimensional cases, and possible to customize kernel functions.

8) Artificial Neural Network Classification (ANN): This class of ANN is one of the simplest types of neural networks

[17]. It is also a fed forward algorithm as it passes information in one direction from input neurons through one or more hidden layers to output neurons. The main advantages of using an ANN classifier are the ability to work with incomplete knowledge, storing information on the entire network, having a distributed memory, and having fault tolerance.

9) Random Forest Classifier (RF): It is a non-linear classification technique, which consists of a group of decision trees. [18]. It integrates multiple decision trees to get more accurate predictions. Each decision tree model is used when employed on its own. This algorithm is called random because they choose predictors randomly at a time of training. In addition, it is called a forest since it takes the result of multiple trees to make a decision. The main advantage of Random forests compared to decision trees is the large number of uncorrelated tree models that work as a single unit will always outperform the individual tree models.

B. Phase 2: Water Quality Index Prediction

The second phase of the proposed framework is responsible for the predictive analysis of the water quality index. In this phase, we examined the impact of the water quality index (WQI) in predicting water quality using six regression models. This analysis started by calculating WQI for the dataset using a mathematical model specified in equations 1, 2, 3, and 4 [41]. After that, six regression models have been applied for predicting water quality. These models are LGBM regression, XGB regression, Extra Trees regression, Decision Tree Regression, Random Forest regression, and linear regression [42]. Fig. 2 explains how the six regression models are applied to predict the water quality index.

$$K = \frac{1}{\Sigma(\frac{1}{S_i})} \qquad (1)$$

Where, S_i is the standard value for each variable of water elements, and K is a constant.

Then, the weight value Wi of each element can be calculated as in equation 2.

$$Wi = \frac{k}{Si} \qquad (2)$$

The Quality Impact Qi value for each element in the water dataset can be calculated as in equation 3.

$$Qi = 100 * (\frac{observe\ values\ (oi) - initial\ value\ (li)}{standard\ value\ (si) - initial\ value\ (li)}) \qquad (3)$$

Finally, the water quality index WQI can be calculated as in equation 4.

$$WQI = \sum_{i=1}^{N} Wi * Qi \qquad (4)$$

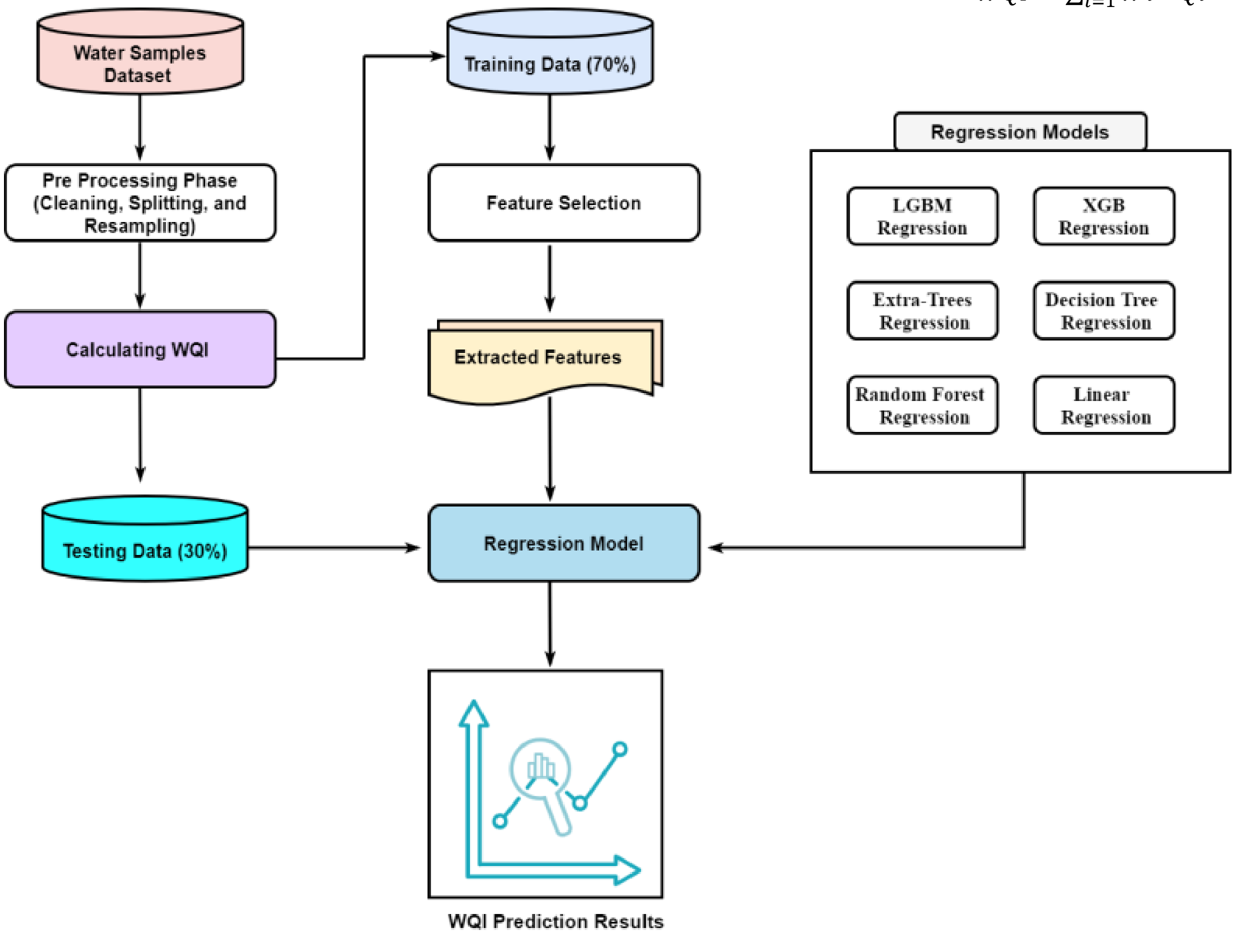


Fig. 2. Water quality index prediction model.

IV. Experimental Results and Discussion

In this section, we present two types of analysis for investigating the efficiency of the proposed machine learning approach in predicting water quality. Subsection A discusses the classification analysis of water samples using nine classifiers, while subsection B discusses the predictive analysis using five regression models:

A. Classification Analysis

The first set of analyses examined the efficiency and accuracy of nine machine learning models used in the proposed framework (as explained in section 3.1) for classifying water samples to recognize that good samples are suitable for human drinking. These performances of these models have been applied to a dataset called Water quality, which was downloaded from [19]. The used dataset consists of 7996 samples of water and 19 features (i.e. variables) that impact water quality. The data has been segmented into training data (6396 samples, 19 features), and testing data (1600 samples, 19 features). The main objective was to classify water samples as suitable for human drinking or not suitable for human drinking. The performance of the nine machine learning models used in the proposed framework has been tested and evaluated using twelve measures as detailed in Table I. The best performance among the nine machine learning models according to each measure is being highlighted. The obtained results clarify that although the random forest algorithm achieved the best training accuracy, the Light GBM outperformed the other classifiers in recognizing good water samples regarding testing accuracy, sensitivity, AUC, F1-score, recall, precision, and mean square error. Fig. 3 and 4 present the comparison results of classification analysis metrics and mean square error (MSE) to nine classifiers, respectively. In addition, Fig. 5 to 13 depicts the performance matrices (or confusion matrices) and the corresponding receiver operating characteristic (ROC) curves of nine machine-learning models, respectively.

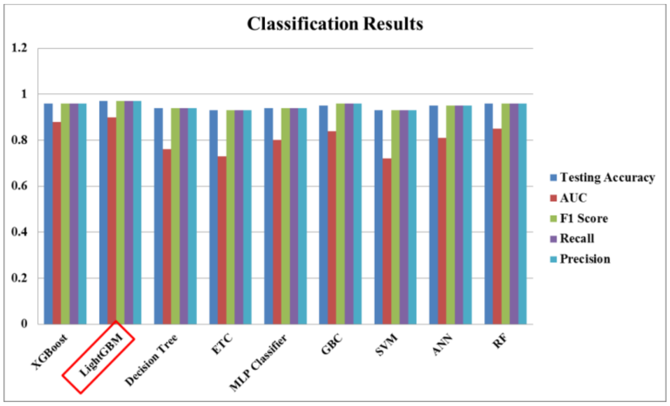


Fig. 3. Comparison results of classification analysis to nine classifiers.

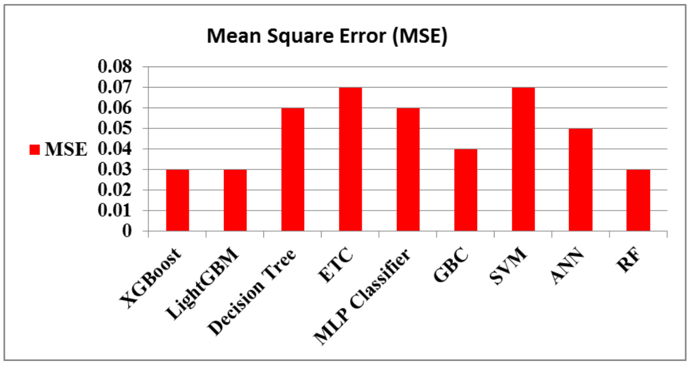
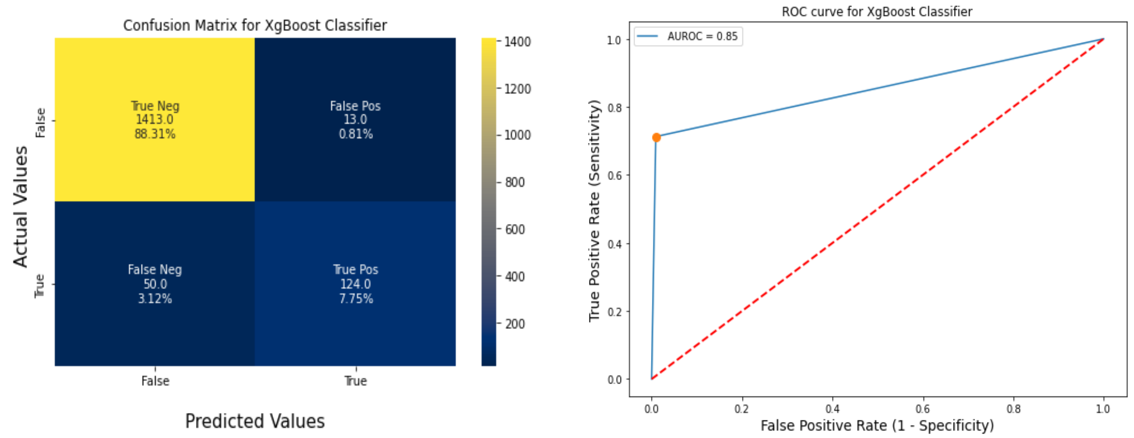


Fig. 4. Comparison of results of Mean Square Error (MSE) to nine classifiers.

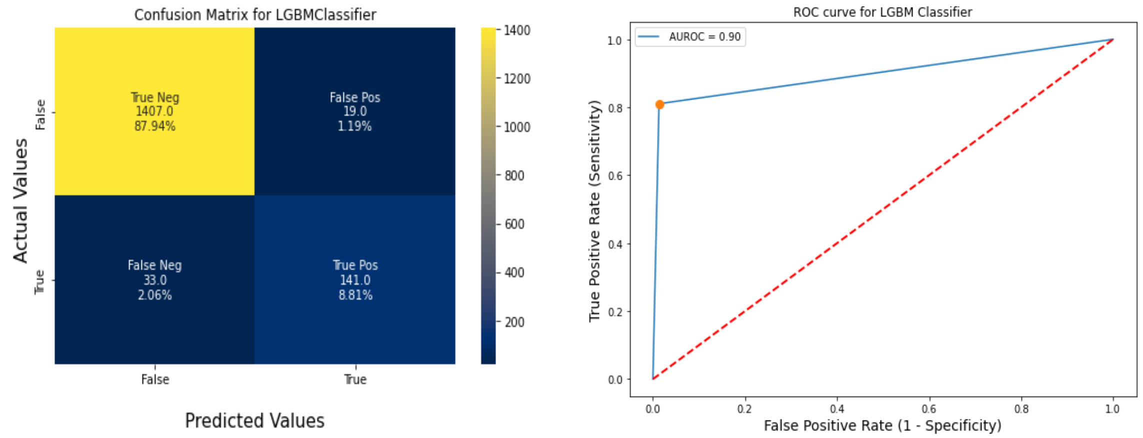
TABLE I. Performance Evaluation Results of Nine Machine Learning Models used in the Proposed Framework

Measure	XGBoost	LightGBM	Decision Tree	ETC	MLP Classifier	GBC	SVM	ANN	RF
Training accuracy	0.97	0.99	0.94	1.0	0.97	0.97	0.95	0.94	1.0
Testing Accuracy	0.96	0.97	0.94	0.93	0.94	0.95	0.93	0.95	0.96
Sensitivity	0.97	0.98	0.94	0.93	0.95	0.96	0.94	0.96	0.97
Specificity	0.91	0.88	0.88	0.88	0.8	0.91	0.82	0.89	0.94
NPV	0.71	0.81	0.52	0.45	0.62	0.70	0.45	0.63	0.71
AUC	0.88	0.90	0.76	0.73	0.80	0.84	0.72	0.81	0.85
F1 Score	0.96	0.97	0.94	0.93	0.94	0.96	0.93	0.95	0.96
Recall	0.96	0.97	0.94	0.93	0.94	0.96	0.93	0.95	0.96
Precision	0.96	0.97	0.94	0.93	0.94	0.96	0.93	0.95	0.96
Mean SE	0.03	0.03	0.06	0.07	0.06	0.04	0.07	0.05	0.03



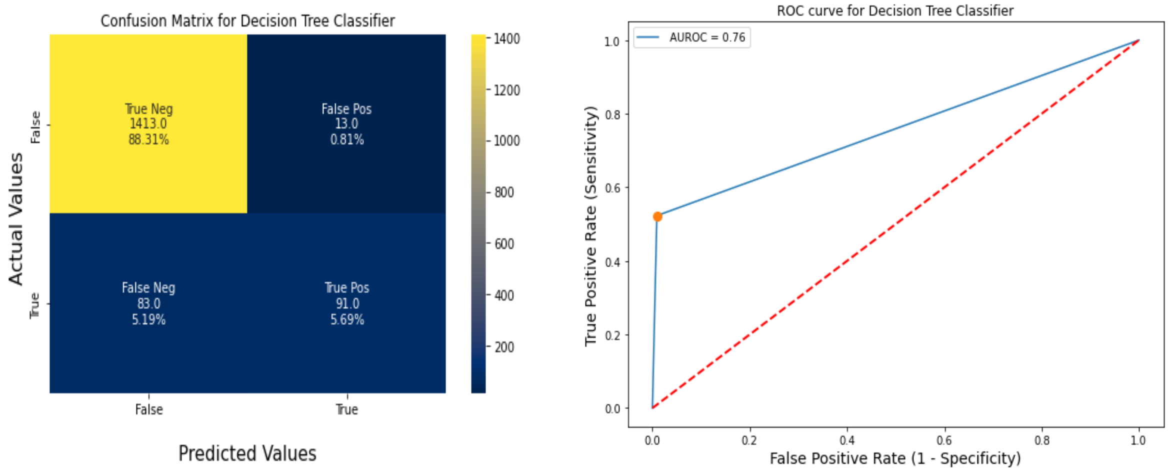
(a) Performance Matrix. (b) FPR vs TPR curve.

Fig. 5. (a) Performance Matrix and (b) FPR vs TPR curve of XGBoost classifier.



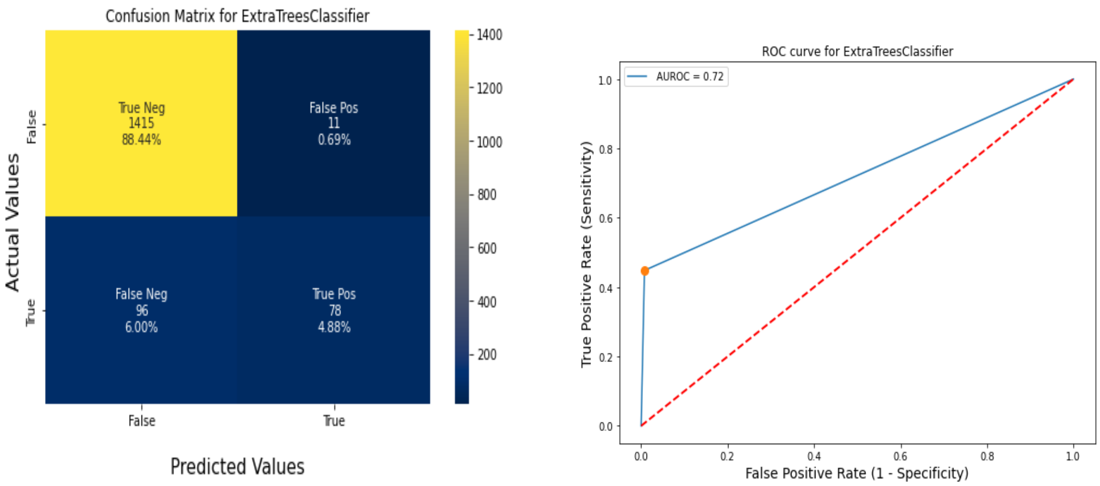
(a) Performance Matrix. (b) FPR vs TPR curve.

Fig. 6. (a) Performance Matrix and (b) FPR vs TPR curve of LightGBM classifier.



(a) Performance matrix. (b) FPR vs TPR curve.

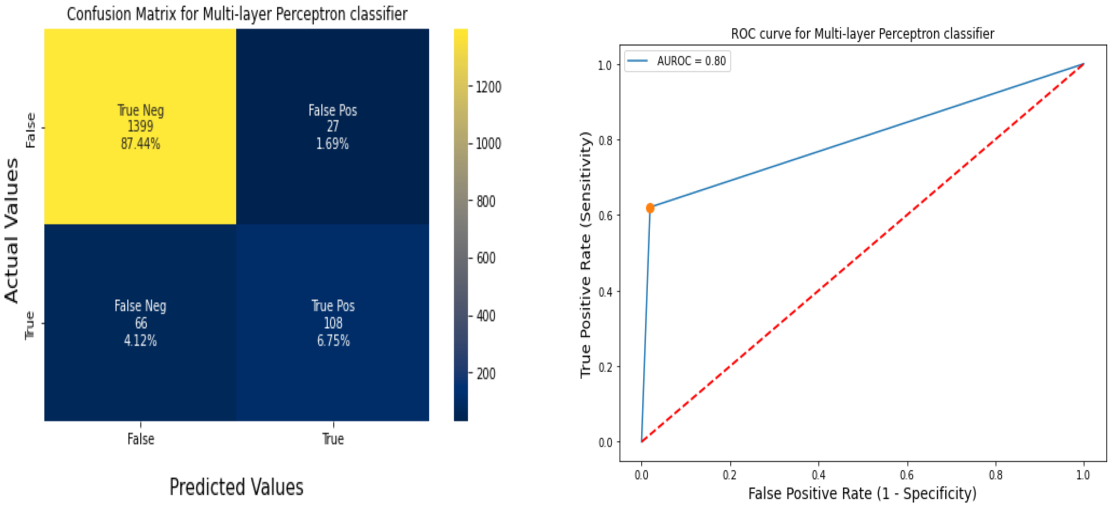
Fig. 7. (a) Performance Matrix and (b) FPR vs TPR curve of Decision Tree classifier.



(a) Performance Matrix.

(b) FPR vs TPR curve.

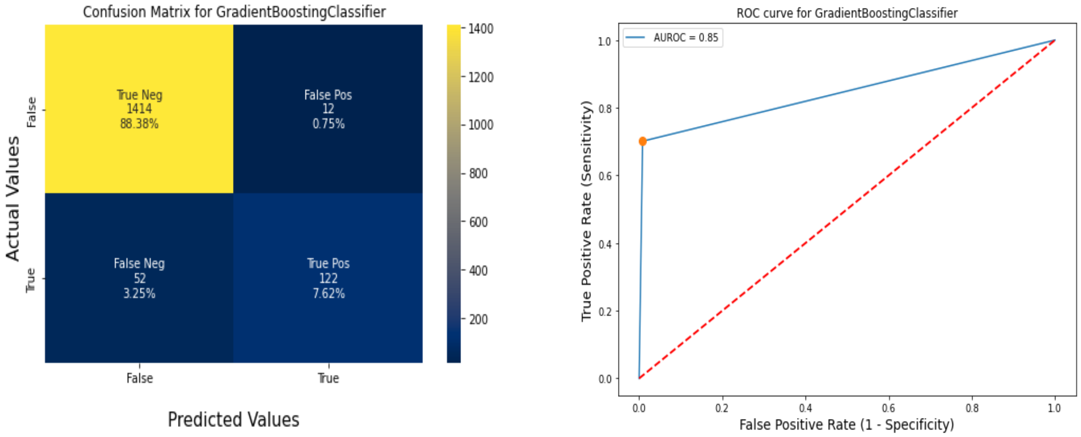
Fig. 8. (a) Performance Matrix and (b) FPR vs TPR curve of Extra Trees Classifier (GBC).



(a) Performance Matrix.

(b) FPR vs TPR curve.

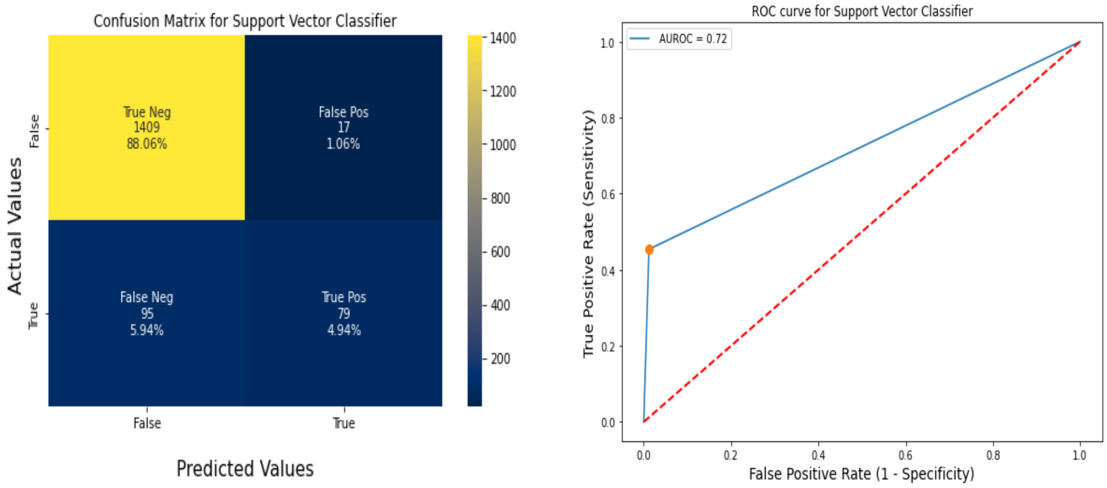
Fig. 9. (a) Performance Matrix and b) FPR vs TPR curve of MLP Classifier.



(a) Performance Matrix.

(b) FPR vs TPR curve.

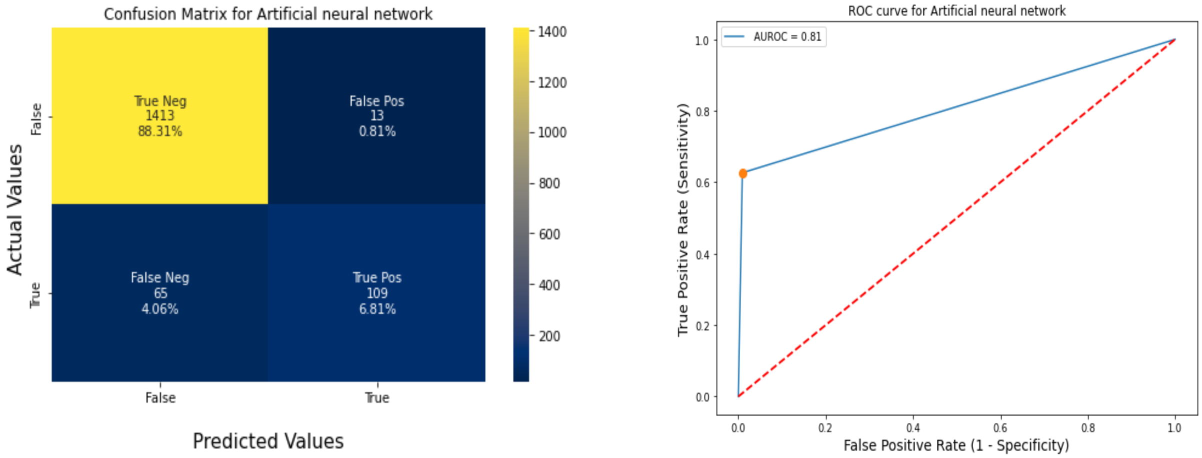
Fig. 10. (a) Performance Matrix and b) FPR vs TPR curve of GB Classifier.



(a) Performance Matrix.

(b) FPR vs TPR curve.

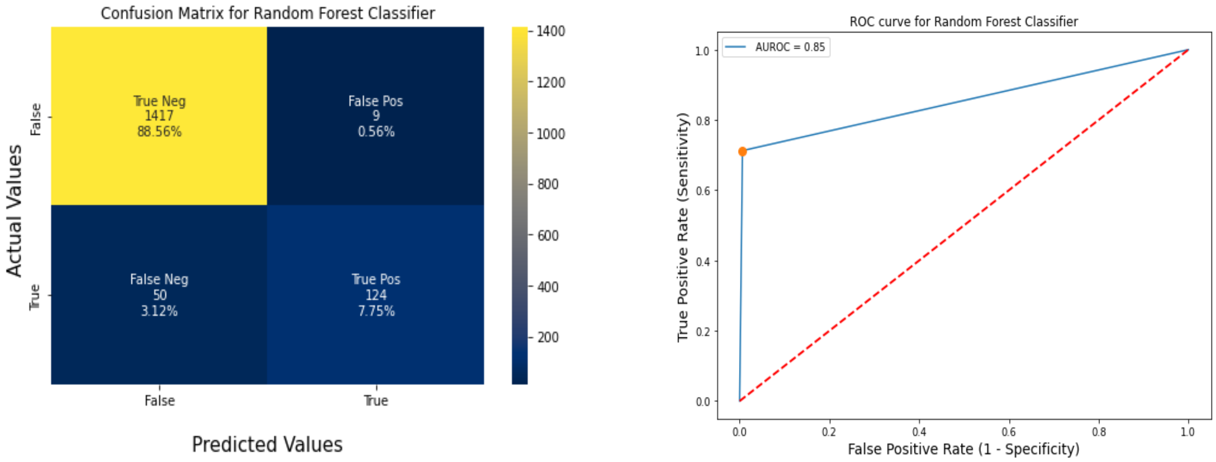
Fig. 11. (a) Performance Matrix and b) FPR vs TPR curve of SVM Classifier.



(a) Performance Matrix.

(b) FPR vs TPR curve.

Fig. 12. (a) Performance Matrix and (b) FPR vs TPR curve of ANN Classifier.



(a) Performance Matrix.

(b) FPR vs TPR curve.

Fig. 13. (a) Performance Matrix and (b) FPR vs TPR curve of RF Classifier.

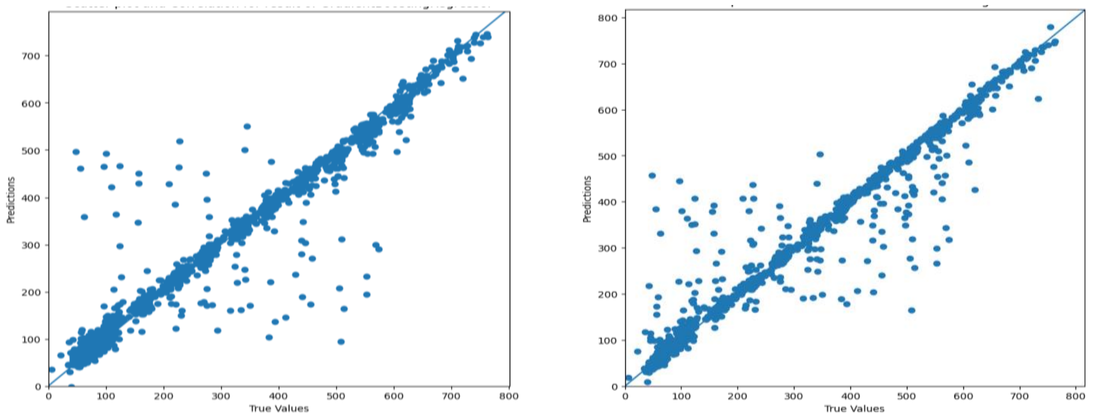
B. *Predictive Analysis*

The second set of analyses examined the efficiency and accuracy of six regression machine learning models used in the proposed framework (as explained in section 3.2) for predicting WQI. Table II summarizes the predictive analysis results of the six regression models after applying the mathematical model of WQI in the dataset. The obtained results have been evaluated based on the common regression metrics, training accuracy, testing accuracy, R2, Adjusted R2, and Mean absolute error (MAE).

The regression analysis results show the superiority of LGBM regression, and Extra Trees Regression models in predicting water quality index according to training and testing accuracy as well as the mean absolute error (MAE) compared to the other regression models. Fig. 14 to 16 visualizes the prediction results of the six regression models, respectively. Fig. 17 presents the comparison results of regression analysis of the used six regression models.

TABLE II. REGRESSION ANALYSIS RESULTS

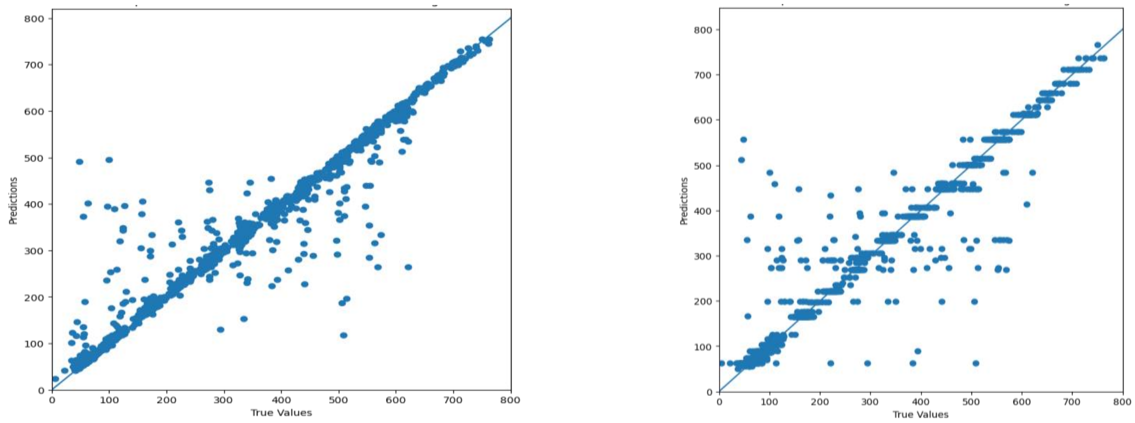
Models and Measurement	**LGBM Regression**	**XGB Regression**	**Extra Trees Regression**	**Decision Tree Regression**	**Random Forest Regression**	**Linear Regression**
Training Accuracy	99.0	97.6	99.9	95.4	98.8	92.9
Testing Accuracy	95.5	95.4	95.5	94	94.8	93.5
$\boldsymbol{R^2}$	94.2	95.47	95.55	94.09	90.6	90.6
Adjusted $\boldsymbol{R^2}$	94.1	95.41	95.2	94.02	90.5	90.5
MAE	10.88%	15.754%	10.07%	17.45%	15.35%	19.34%



(a) LGBM Regression.

(b) XGB Regression.

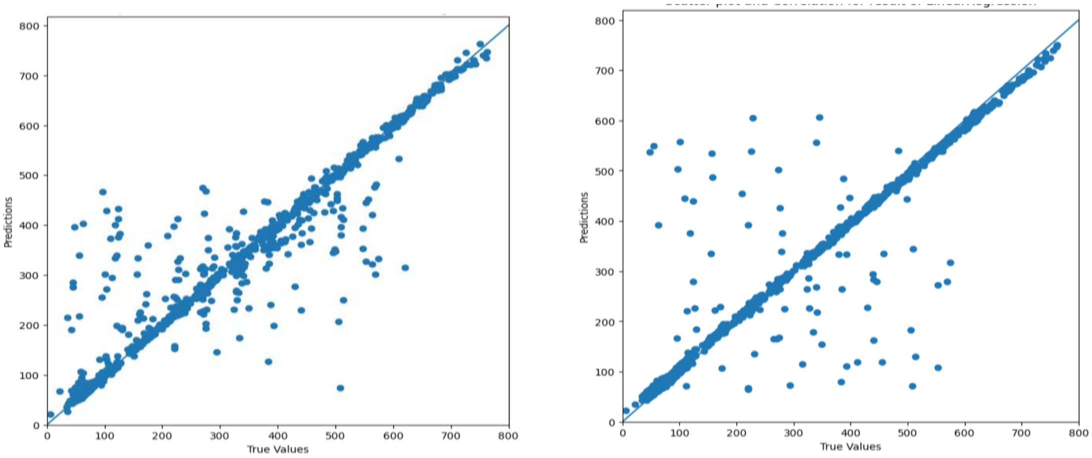
Fig. 14. Regression analysis results of a) LGBM Regression and b) XGB Regression.



(a) ET Regression.

(b) DT Regression.

Fig. 15. Regression analysis results of a) ET Regression and b) DT Regression.



(a) RF Regression.

(b) Linear Regression.

Fig. 16. Regression analysis results of a) RF regression and b) Linear regression.

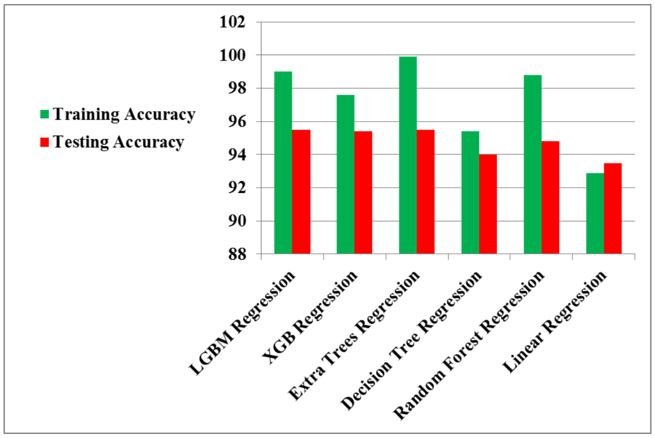


Fig. 17. Comparison results of regression analysis results of the used six regression models.

V. Conclusion

The present article was designed to investigate the efficiency of using a proposed machine-learning framework to classify drinking water samples and predict water quality index. The classification tier of the proposed framework consists of nine classification models, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting (LightGB), Decision Tree (DT), Extra Tree (ET) classifier, Multi-layer Perceptron (MLP) classifier, the Gradient Boosting (GB) classifier, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest (RF) classifier. The performance of those models has been validated on a benchmark dataset consisting of 7996 water samples, and 19 features. The obtained results clarified good classification results to the nine models with average accuracy. 94.7%. However, the obtained results clarified that, although the Random Forest (RF) algorithm achieved the best training accuracy, 100%, the Light GBM outperformed the other classifiers in recognizing good water samples regarding testing accuracy, 0.97%. The second goal of this study was to investigate the efficiency of the regression tier through applying six regression models for predicting water quality index. The regression analysis clarified the superiority of LGB regression, and Extra Trees Regression models in predicting water quality index according to training and testing accuracy as well as the mean absolute error (MAE) compared to the other regression models. Taken together, these findings suggest a role for using machine learning models in promoting the analysis and prediction of water quality. Moreover, these results have significant implications for the understanding of how novel deep learning models can be developed for predicting water quality, which is suitable for human drinking, irrigation of plants and crops, and other industrial or environmental purposes.

References

[1] Wolfram J, Stehle S, Bub S, Petschick LL, Schulz R. Water quality and ecological risks in European surface waters–Monitoring improves while water quality decreases. Environment International. 2021 Jul 1;152:106479.

[2] World Health Organization. A global overview of national regulations and standards for drinking-water quality. [online], available at: https://apps.who.int/iris/bitstream/handle/10665/350981/9789240023642-eng.pdf?sequence=1 (accessed 19/4/2022).

[3] United Nations, Water [online], available at: https://www.un.org/en/global-issues/water (Accessed 19/4/2022).

[4] Alamanos A, Mylopoulos N, Loukas A, Gaitanaros D. An integrated multicriteria analysis tool for evaluating water resource management strategies. Water. 2018 Dec;10(12):1795.

[5] Schmutz S, Sendzimir J. Riverine ecosystem management: Science for governing towards a sustainable future. Springer Nature; 2018.

[6] Tung TM, Yaseen ZM. A survey on river water quality modeling using artificial intelligence models: 2000–2020. Journal of Hydrology. 2020 Jun 1;585:124670.

[7] Hmoud Al-Adhaileh M, Waselallah Alsaade F. Modelling and prediction of water quality by using artificial intelligence. Sustainability. 2021 Jan;13(8):4259.

[8] Hameed M, Sharqi SS, Yaseen ZM, Afan HA, Hussain A, Elshafie A. Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in the tropical region, Malaysia. Neural Computing and Applications. 2017 Dec;28(1):893-905.

[9] Chen Y, Song L, Liu Y, Yang L, Li D. A review of the artificial neural network models for water quality prediction. Applied Sciences. 2020 Jan;10(17):5776.

[10] Osman AI, Ahmed AN, Chow MF, Huang YF, El-Shafie A. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal. 2021 Jun 1;12(2):1545-56.

[11] Gan M, Pan S, Chen Y, Cheng C, Pan H, Zhu X. Application of the machine learning LightGBM model to the prediction of the water levels of the lower Columbia River. Journal of Marine Science and Engineering. 2021 May;9(5):496.

[12] Lu H, Ma X. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere. 2020 Jun 1;249:126169.

[13] Upadhyay R, Tanwar PS, Degadwala S. Fracture Type Identification Using Extra Tree Classifier. In2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud)(I-SMAC) 2021 Nov 11 (pp. 1-6). IEEE.

[14] Haribabu S, Gupta GS, Kumar PN, Rajendran PS. Prediction of Flood by Rainf All Using MLP Classifier of Neural Network Model. In2021 6th International Conference on Communication and Electronics Systems (ICCES) 2021 Jul 8 (pp. 1360-1365). IEEE.

[15] Khan MS, Islam N, Uddin J, Islam S, Nasir MK. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University-Computer and Information Sciences. 2021 Jun 14.

[16] Singh KP, Basant N, Gupta S. Support vector machines in water quality management. Analytica chimica acta. 2011 Oct 10;703(2):152-62.

[17] Sulaiman K, Ismail LH, Razi MA, Adnan MS, Ghazali R. Water quality classification using an Artificial Neural Network (ANN). In IOP Conference Series: Materials Science and Engineering 2019 Aug 1 (Vol. 601, No. 1, p. 012005). IOP Publishing.

[18] Ko BC, Kim HH, Nam JY. Classification of potential water bodies using Landsat 8 OLI and a combination of two boosted random forest classifiers. Sensors. 2015 Jun;15(6):13763-77.

[19] MsS.Pants, Water quality, [online], available at https://www.kaggle.com/datasets/mssmartypants/water-quality.

[20] Mustafa HM, Mustapha A, Hayder G, Salisu A. Applications of IoT and Artificial Intelligence in Water Quality Monitoring and Prediction: A Review. In 2020 6th International Conference on Inventive Computation Technologies (ICICT) 2021 Jan 20 (pp. 968-975). IEEE.

[21] Rajaee T, Khani S, Ravansalar M. Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review. Chemometrics and Intelligent Laboratory Systems. 2020 May 15;200:103978.

[22] Ahmed AN, Othman FB, Afan HA, Ibrahim RK, Fai CM, Hossain MS, Ehteram M, Elshafie A. Machine learning methods for better water quality prediction. Journal of Hydrology. 2019 Nov 1;578:124084.

[23] Haghiabi AH, Nasrolahi AH, Parsaie A. Water quality prediction using machine learning methods. Water Quality Research Journal. 2018 Feb;53(1):3-13.

[24] Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, García-Nieto J. Efficient water quality prediction using supervised machine learning. Water. 2019 Oct 24;11(11):2210.

[25] Abubakr Yahya AS, Ahmed AN, Binti Othman F, Ibrahim RK, Afan HA, El-Shafie A, Fai CM, Hossain MS, Ehteram M, Elshafie A. Water quality prediction model based support vector machine model for ungauged river catchment under dual scenarios. Water. 2019 Jun 13;11(6):1231.

[26] Azrour M, Mabrouki J, Fattah G, Guezzaz A, Aziz F. Machine learning algorithms for efficient water quality prediction. Modeling Earth Systems and Environment. 2022 Jun;8(2):2793-801.

[27] Aldhyani TH, Al-Yaari M, Alkahtani H, Maashi M. Water quality prediction using artificial intelligence algorithms. Applied Bionics and Biomechanics. 2020 Dec 30;2020.

[28] Nafi SN, Mustapha A, Mostafa SA, Khaleefah SH, Razali MN. Experimenting with two machine learning methods in classifying river water quality. International Conference on Applied Computing to Support Industry: Innovation and Technology 2019 Sep 15 (pp. 213-222). Springer, Cham.

[29] Wu J, Wang Z. A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory. Water. 2022 Feb 17;14(4):610.

[30] Than NH, Ly CD, Van Tat P. The performance of classification and forecasting Dong Nai River water quality for sustainable water resources management using neural network techniques. Journal of Hydrology. 2021 May 1;596:126099.

[31] Yan J, Liu J, Yu Y, Xu H. Water quality prediction in the luan river based on 1-DRCNN and bigru hybrid neural network model. Water. 2021 Apr 30;13(9):1273.

[32] Sha J, Li X, Zhang M, Wang ZL. Comparison of forecasting models for real-time monitoring of water quality parameters based on hybrid deep learning neural networks. Water. 2021 May 31;13(11):1547.

[33] J.Brownlee. Extreme Gradient Boosting (XGBoost) Ensemble in Python, Machine Learning Mastery (27/4/2021), [online], available: https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/ (accessed 22/4/2022).

[34] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd ACM signed international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

[35] J.Brownlee , How to Develop a Light Gradient Boosted Machine (LightGBM) Ensemble, Machine Learning Mastery (25/11/2020), [online], available at: https://machinelearningmastery.com/light-gradient-boosted-machine-lightgbm-ensemble/ (accessed 22/4/2022).

[36] Learn website, Decision Trees [online], available at: https://scikit-learn.org/stable/modules/tree.html#:~:text=Decision%20Trees%20(DTs)%20are%20a,as%20a%20piecewise%20constant%20approximation. (accessed 23/4/2022).

[37] A.Gupta, ML | Extra Tree Classifier for Feature Selection, [online], available at : https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/ (accessed 23/4/2022).

[38] Swarnimrai, Multi-Layer Perceptron Learning in Tensorflow, [online], available: https://www.geeksforgeeks.org/multi-layer-perceptron-learning-in-tensorflow/?ref=gcse (accessed 23/4/2022).

[39] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in neurorobotics. 2013 Dec 4;7:21. [online], available at: https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021/full (accessed 10/5/2022).

[40] Geeks for Geeks website, Support Vector Machine Algorithm, [online], available at: https://www.geeksforgeeks.org/support-vector-machine-algorithm/?ref=gcse (accessed 10/5/2022).

[41] Kadam AK, Wagh VM, Muley AA, Umrikar BN, Sankhua RN. Prediction of water quality index using artificial neural network and multiple linear regression modeling approach in Shivganga River basin, India. Modeling Earth Systems and Environment. 2019 Sep;5(3):951-62.

[42] Maulud D, Abdulazeez AM. A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends. 2020 Dec 31;1(4):140-7.