Applied Mathematics & Information Sciences
*An International Journal*

http://dx.doi.org/10.18576/amis/190620

# Comparative Analysis of River Water Quality in Asia and the Arab Region Using Machine Learning-Based Classification and Prediction Models

*Hassan Shaheed[1,2,*], Mohd Hafiz Zawawi[1], Gasim Hayder[1], Karim Sherif Mostafa[3], Norbaya Sidek[4], Mohamed A. Hafez[5,6].*

[1] Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Kajang 43000, Selangor Darul Ehsan, Malaysia
[2] Ministry of Planning of Republic of Iraq, Baghdad P.O. Box 13032, Iraq
[3] Civil Engineering Discipline, School of Engineering, Monash University, Malaysia
[4] Faculty of Civil Engineering, Universiti Teknologi MARA (UiTM), Selangor, Malaysia
[5] Faculty of Engineering and Quantity Surveying INTI -IU, Universi, Nilai -Malaysia
[6] Faculty of Management, Shinawatra University, Pathum Thani Thailand

Received: 2 Jul. 2025, Revised: 12 Sep. 2025, Accepted: 12. Oct. 2025
Published online: 1 Nov. 2025

**Abstract:** Water quality is a key pillar of environmental stability and human health, but it is increasingly threatened by a variety of contaminants. These pollutants disturb water balance and pose serious public health problems. As the challenges of water quality increase, the simulation and forecasting of water quality are becoming essential to address contamination problems. Recent research has focused on the development of comprehensive models capable of classifying and predicting water quality using advanced machine-learning algorithms to classify water quality index (WQI0 values according to predefined parameters. Synthetic pollution index (SPI) and the water quality index (WQI) are among the most often used methods for classifying and reflecting the quality of the water and pollution risk in a given area [46]. Data were gathered from local rivers in Malaysia, Iraq, and India. The WQI was calculated using 32 parameters, including temperature, dissolved oxygen, hardness, pH, coliforms, and chloride concentrations. Pre-processing of the data involved class imbalance, outliers, and standardization. An automated water quality assessment system has been developed using a hybrid approach combining CatBoost, Support Vector Machine (SVM), Naive Bayes, and Light-Blight Gradient (LGBM) regression models with Random Forest (RF), EML Regressor, Decision Tree (DT), and M5 Model Tree (MLM) regression. CatBoost achieved the highest classification accuracy of 94.55 percent, with a Matthews correlation coefficient (MCC) of 93.31 percent for the Malaysian dataset. For regression analysis, the M5 model tree had a superior predictive performance for most datasets and had strong results for the metrics MAE, MSE, and $R^2$, while Random Forest had better results for the Malaysian dataset. The results highlight the spatial diversity of water quality in the study regions and confirm the ability of machine learning models to support water quality management based on the data. The proposed models showed high accuracy and reliability in the WQI classification.

**Keywords:** Classification Models; River Water Quality; Machine Learning Algorithms; Water Quality Index.

## 1. Introduction

Although water covers about 71 percent of the surface of the Earth, only about three percent of it is fresh water, and less than one percent of that fraction is available for human consumption. By 2025, almost 40 per cent of the world's population is expected to live in freshwater-scarce areas [1]. Rapid population growth and continued economic expansion are among the main drivers of the increasing demand for freshwater. Rising pollution levels further reduce the availability of drinking water. Reports by the United Nations, the World Health Organisation, and UNICEF show that more than 2.2 billion people do not have access to safe drinking water. Studies in developing countries show that about 80 percent of health problems are linked to water contamination. An estimated five million people die every year from water-borne diseases affecting about 2.5 billion people worldwide [2].

Poor management of municipal waste and hazardous chemicals can pollute rivers, lakes, and groundwater and lead to deterioration of both the quality of surface and groundwater [3], [4]. Therefore, continuous monitoring of water quality is necessary to detect possible pollutant ingress. The increasing variety and complexity of contaminants have made manual detection methods inadequate to cope with modern pollution problems. Considering the proven reliability and accuracy of smart computing approaches in many scientific and engineering fields, their use in water quality monitoring is expected to increase. Machine learning (ML) and deep learning (DL) techniques are increasingly being adopted in a wide range

*Corresponding author e-mail: hassan.en37@yahoo.com

of scientific and engineering fields, including cyber security [5], agriculture [6], environmental monitoring [7], sentiment analysis [8], and medicine [9]. Since the advent of ML, there has been steady progress in the application of artificial intelligence for the prediction of surface water quality [11], [12]. However, many countries still face constraints, particularly the insufficient establishment of large-scale monitoring systems for water quality. These constraints limit the availability of the real-time data required for accurate evaluation and forecasting. Integrating ML and DL approaches offers a promising way to overcome these limitations by enabling automated data-driven analysis, which improves both forecasting accuracy and management effectiveness.

The accuracy of WQI analysis and forecasting requires the use of innovative calculation methods. Studies dealing with the time dimension of water quality forecasting are encouraged to include seasonal variations in the modelling of WQI trends. Employing tailored model variations often yields a higher predictive accuracy than relying on one algorithm alone. Multiple analytical approaches, including algorithmic analysis, visual modelling, and statistical evaluation[13], can be used to classify water quality based on the WQI [14]. In this context, reinforcement algorithms have demonstrated strong performance across a wide range of datasets from several countries. Among these, the CatBoost and Light-Bolt Generation Machine (LGBM) achieved higher accuracy when applied in the experimental framework described in [15]. In Malaysia, water quality assessment and hydrological analysis often rely on six key parameters: pH, chemical oxygen demand (COD), dissolved oxygen (DO), ammonium nitrate (NH-N), bicarbonate (BC), and dissolved solids (SS). The WQI used in these studies was formulated without the inclusion of dangerous chemicals or microbiological indicators such as microorganisms, heavy metals, or pesticides, so that the WQI reflected only the general physicochemical quality of the water and not the presence of toxic or biological contaminants.

In Malaysia, the Ministry of Environment monitors 23 water quality parameters; however, but only six of these indicators are used for the overall WQI. Given the growing threat of water resource contamination, comprehensive monitoring of multiple parameters is essential to prevent outbreaks of waterborne diseases. This study makes two main contributions. First, it assesses the performance of a few conventional machine learning classification engines using a variety of datasets from Malaysia, India, and Iraq. Second, it integrates reinforcement algorithms with synthetic minority oversampling techniques (SMOTE) to improve the accuracy of the model and address the class imbalance in water quality predictions. For the Indian dataset, seven sampling locations collected water quality data from the Bhavani River, which straddles Tamil Nadu and Kerala. For the Malaysian dataset, Klang and Langat River data from two monitoring stations were collected. Information from Iraq for the Tigris and Euphrates Rivers was collected from six sampling points. Machine learning models have been used to classify water quality indicators using metrics such as recall, accuracy, precision, F1 scores, and the Matthews correlation coefficient. In addition, water quality was assessed by regression analysis using a decision tree regressor, M5 model tree, random forest regressor, and extreme learning machine. The regression models were evaluated using performance metrics including mean squared error (MSE), mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), scatter index (SI), and bias.

## 2. Related Work

This study aimed to identify effective methodological approaches to address water quality problems. While conventional laboratory analysis and statistical methods remain the standard tools for water quality assessment, this study explored the potential of machine learning algorithms to improve the accuracy of the assessment and support the development of more robust solutions to water quality challenges. Sillberg et al.. (2021) [16] introduced the use of machine learning methodology combining SVM and AR for the classification of the water quality of the Chao Phraya River. The model achieved F1 values of 0.84, mean 0.84, accuracy 0.84, and error 0.94. The best classification results were observed using a linear SVM. The use of SVM to demonstrate the proposed method shows significant improvements in WQI and achieves accuracies of 0.86 to 0.95 using six relevant parameters.

Yilma and colleagues. (2018) [17] reconstructed the Akaki River WQI using an artificial neural network (ANN). The model included 12 water quality indicators collected from 27 locations under varying temperature and precipitation conditions. An ANN structure consisting of eight hidden layers and 15 neurons achieved a WQI of 0.93. In a different study, Ahmed and Others. (2019) [18] measured the WQI using supervised machine learning algorithms, in which the total water quality and its classification are represented by a single index. The WQI achieved a classification accuracy of 85.07 percent. Among the algorithms tested, gradient boosting with a learning rate of 0.1 and second-degree polynomial regression showed the highest predictive power. Sakizadeh [19] used 16 water quality parameters combined with a Bayesian regularization approach to calculate the WQI. The correlation coefficients between the actual and forecast values were 0.77 and 0.94, which indicated the robustness of the model. This study builds on previous research by demonstrating the ability of machine learning as a reliable approach for detecting water quality anomalies. In addition, the use of machine learning techniques is expected to significantly reduce the probability of false predictions. Several studies have used multilevel modelling techniques and frameworks to assess water quality [20], [21].

Research in this area is increasingly focused on integrating machine learning and big data (BD) to provide a comprehensive analysis of the factors affecting WQI in rivers around the world. In [14], predictive models for the classification of river water quality were developed using ML index modelling. The study used several algorithms, including SVM, Naive Bayes, DT, and MLP, with a particular focus on the Bhavani River. The developed model achieved a classification accuracy exceeding 0.81. In a previous study [22], researchers combined several predictive models, including RNN, LSTM, RF, SVR, MLP, and linear regression. Among these models, LSTM achieved the highest performance, with an accuracy of more than 88 per cent. A further study [23] used several predictive approaches, linear regression, LSTM, Gated Recurrence Unit (GRU), RNN, RF, SVR, and MLP, to analyze the dataset of the Bhavani River, where the temporal fusion transform (TFT) was shown to have a higher predictive power. A recent study [24] also adopted the Adam Optimizer TFT architecture and reported a consistent performance with other models. Further research [25], [26], [27], and [28] used different machine learning algorithms to predict and classify WQI. Despite these advances, ensemble algorithms such as the XGBoost family have only recently been studied [29],[30], [31]. On the other hand, a separate study [32] estimated the WQI for Iraq using conventional statistical methods but without incorporating machine learning models.

Ann and colleagues. (2021) [35] used the long-term short-term memory (LSTM NN) network to identify time patterns in water quality data and to predict the future values of parameters. For the collection of real-time data, pH, turbidity, and total dissolved solids (TDS) sensors were used. The system relied on an Internet of Things (IoT) module installed in the water supply, integrating sensors with an Arduino and a NodeMCU for regular monitoring. The main advantage of this approach is its ability to provide early warnings of potential contamination, allowing preventive actions to be taken before water becomes unsafe. Several researchers have also proposed a framework for water quality assessment and forecasting using machine learning techniques [36]. Prior to training the model, the datasets underwent preprocessing steps that included data cleaning, partitioning, correlation analysis, and standardization of the Z-score. The water quality class (WQC) was predicted by means of SVM, XGBoost, decision tree, and random forest models, and the WQI was measured by means of linear regression models such as LSTM, SVR, MLP, and NARNet. Two datasets were used to evaluate the model: Dataset 1 contained 600 data points per parameter and Dataset 2 contained 6,000 data points per parameter. Differences in the dataset size allowed a direct comparison of the performance of the models. The MLP regression model demonstrated superior prediction capability, as reflected in lower MAE, MSE, and RMSE values, with the best fit of 0.93 $R^2$.

Mehreen [37] used several regression models, including the LightGBM, MLP, and SVM models, which used air pollutants and meteorological conditions as predictor variables. RMSE ranged between 0.015 and 0.18 for all models. MLP produced RMSE values of 0.18 for TDS and 0.003 for pH, while SVM achieved RMSE values ranging from 0.015 to 0.027 for DO, turbidity, and ECC. In a follow-up study, Mehreen [38] proposed a refined water quality assessment framework based on a semi-supervised machine learning algorithm to design a Water Quality Index. The process involved the selection of parameters, calculation of sub-indexes, weighting, aggregation of sub-indices, and classification. Data on the Rawalc river basin network were collected and covered physicochemical, atmospheric, meteorological, hydrological, and topographical variables. The sub-indices were normalized by the min-max method and weight was determined using tree algorithms such as LightGBM, Random Forest, CatBoost, AdaBoost, and XGBoost. The proposed method achieves a classification rate of 100 percent and removes the uncertainties associated with traditional indexing by eliminating the requirement that all classification parameters are included. The parameters most affected were electrical conductivity, Secchi disk depth, dissolved oxygen, lithology, and geology. LightGBM achieved 99.1 percent accuracy and CatBoost 99.3 percent. Mustafa [39] used XGBoost, AdaBoost and Random Forest (RF) algorithms to classify binary data and used random search and grid searches to optimize hyperparameters. The hybrid model combining support vector regression and XGBoost showed the highest classification performance, with 99.4 percent accuracy and the highest F1 score of all tested models. The hybrid SXH approach outperformed the comparator methods in terms of both precision and F1 metrics.

Peer-reviewed studies highlight the considerable potential of ML algorithms to solve water quality problems. Models such as CatBoost, LightGBM, and XGBoost have consistently demonstrated superior performance in the handling of complex and diverse datasets. Techniques such as SVM and naive Bayes continue to be effective for classification tasks, whereas regression models such as the M5 model tree and random forest show strong predictive power across a variety of conditions. The selection of algorithms in this study is based on their proven ability to resolve class imbalances, capture non-linear relationships, and process high-dimensional data, as supported by peer-reviewed literature.

## 3. Study Area and Dataset Overview

This study used three datasets from India, Malaysia, and Iraq, representing South, Southeast, and West Asia. The distance between Malaysia and Iraq exceeds 6,500 km (Figure 1) and indicates major climatic and hydrological variation. A map from Google Maps shows the main river systems in these regions and underscores the importance of evaluating the water quality across Asia. The datasets

include the Bhavani River in India, Klang and Langat Rivers in Malaysia, and Tigris and Euphrates Rivers in Iraq. Machine learning classifiers were applied to assess the WQI and examine how variations in water quality influence population, agriculture, and environmental conditions.


**Fig. 1:** Map of participating countries

The Bhavani River passes through Tamil Nadu and Kerala. The river originates in the Nilgiri Hills and extends across Tamil Nadu, Kerala, and the Silent Valley Wildlife Sanctuary. Its course includes the Attappady Plateau in the Palakkad district, before it reaches the plains of Tamil Nadu. Although several monitoring stations were located along the Bhavani River, data were collected from only seven stations. The Langat and Klang Rivers run through Kuala Lumpur and Selangor and continue toward Malacca. The Tigris and Euphrates Rivers cross Iraq from north to south and reflect a continuous hydrological connection across Western Asia.

### 3.1 Data Collection

The data analyzed were from rivers in India, Malaysia, and Iraq. The following analysis compares the information pertaining to each of these rivers, as well as the research approaches taken for the survey.

#### 3.1.1 Bhavani River of India

The Bhavani Basin includes seven key monitoring stations: Thavalam, Karathur, Kottathara, Cheerakuzhi, Chalayur, Badrakaliamman Kovil, and Elachivazhi. The WQI for these sites was calculated based on core parameters such as temperature, pH, DO, turbidity, and chloride concentration. The dataset contained 7,649 water samples encompassing 31 physicochemical properties collected between January 1, 2016, and December 22, 2019. The analysis focuses on comparing the water quality of the Bhavani River across seven locations. Statistical results related to the Bhavani dataset are provided in (Figure 1 and Table 1, supplementary materials).

#### 3.1.2 Klang-Langar Rivers of Malaysia

Datasets pertaining to the rivers of Malaysia were collected for the Langat and Klang Rivers, aimed at important variables associated with the respective basins. A total of 656 data samples consisting of six attributes, which were collected at WQ monitoring stations from January 2005 to August 2016, were aggregated for this study. River water quality data-specific information from various sampling locations is illustrated (Figure 2 and Table 2, Supplementary materials). Statistical analysis of the rivers of Malaysia is also provided. The dataset contains six features with different data types, ranging from floats to integers and objects. Visualizations of the Malaysian dataset, among which a histogram, boxplot, violin plot, and correlation matrix are also provided.

#### 3.1.3 Tigris and Euphrates Rivers of Iraq

The main monitoring stations located in the Tigris and Euphrates River basins are labeled as T27, T28, E8, E11, E16, and E19. These stations were the key parameters that affected the WQI, including pH, temperature, DO, chlorides, and turbidity, which were recorded with absolute diligence. Between 2010 and 2019, 481 samples of data, each consisting of 17 different parameters, were obtained from these monitoring stations. Statistical analyses of the Iraqi rivers are presented (Figure 3 and Table 3, Supplementary materials).

### 3.2 WQI Calculation

The WQI is derived using equation (1) for the sole purpose of assigning numerical values to all parameters that characterize water quality. It allows for an overall water quality assessment by assigning appropriate importance to each parameter individually.

$$WQI = \frac{\sum wiqi}{\sum wi} \quad (1)$$

In this regard, $I$ represents the number of parameters considered in the analysis, where $qi$ represents the relative importance of a given water quality parameter. *The Wi* notation is used to find the relative significance of a parameter, which is needed in the calculation of the WQI. *The qi* value can be obtained using Equation (2).

$$qi = 100 * \frac{vi - vo}{si - vo} \quad (2)$$

In this context, the ideal values for all parameters are set to zero, except for DO, which is set to 14.6 mg/L, and pH, which is set to 7. For each parameter analyzed, *vi* represents the value obtained experimentally, while *vo* denotes the theoretical value. *si* refers to the legally defined standard associated with the water category to which the sample belongs. The weighting factor *wi* is calculated using Equation (3).

$$wi = \frac{K}{si} \quad (3)$$

The constant $K$ is calculated based on equation (4),

$$K = \frac{1}{\sum\left(\frac{1}{si}\right)} \quad (4)$$

To calculate the WQI, each parameter was assigned a specific weight and a permissible limit. The measured parameters include temperature (28 °C, 0.0357), pH (8.5,

0.1765), chemical oxygen demand COD (150 mg/L, 0.0067), turbidity (5 NTU, 0.2), conductivity (20 µS/cm, 0.05), total alkalinity (200 mg/L, 0.005), chloride (250 mg/L, 0.004), phenolphthalein alkalinity (10 mg/L, 0.1), total Kjeldahl nitrogen TKN (100 mg/L, 0.01), calcium hardness (50 mg/L, 0.02), magnesium hardness (100 mg/L, 0.01), sulfate (75 mg/L, 0.0133), sodium (30 mg/L, 0.0333), total suspended solids TSS (200 mg/L, 0.005), phosphate (200 mg/L, 0.005), boron (300 µg/L, 0.0033), potassium (1000 mg/L, 0.001), total dissolved solids TDS (200 mg/L, 0.005), fixed dissolved solids FDS (0.3 mg/L, 3.3333), biochemical oxygen demand BOD (1 mg/L, 1), fluoride (2.5 mg/L, 0.4), dissolved oxygen DO (3 mg/L, 0.3333), nitrate-nitrogen $NO_3$–N (1.5 mg/L, 0.6667), fecal coliform FC (7.5 MPN/100 mL, 0.1333), and total coliform TC (0.503 MPN/100 mL, 1.9881) (Figure 4-Figure 15, Supplementary materials). Table 1 presents the ecological conditions of the water based on the weighted arithmetic WQI method and provides an overview of the water quality status at the studied sites.

**Table 1:** Weighted Arithmetic Standards for Water Quality Index

| Water Index | WQI Range | WQI Class | WQI Quality |
|---|---|---|---|
| **1** | 0-30 | A | Good |
| **2** | 31-60 | B | Moderate |
| **3** | 61-90 | C | Poor |
| **4** | 91-120 | D | Very Poor |
| **5** | >121 | E | Unsuitable/Bad |

Weighted arithmetic WQI parameters were used to compute the WQI based on the provided formulas. Each parameter's allowable values and unit weights, as given above, were used to determine the WQI. The WQI value for each sample was determined and used to assign the value to the corresponding instance. The developed criteria were utilized to determine the WQI category of each case, which was used as a class label.

The Bhavani River dataset comprises 33 individual parameters and 7,649 labeled records collected from seven monitoring stations. In comparison, the combined dataset for the Klang and Langat. The Malaysian rivers contain six parameters and a total of 656 labeled records obtained from two monitoring stations. The dataset for the Tigris and Euphrates Rivers of Iraq consists of 17 parameters and 481 labeled records gathered from six monitoring sites. The variation in the number of parameters and records among these datasets provides a broad representation of regional water characteristics and supports a comparative assessment of water quality across different climatic and geographical conditions.

### 3.3 Data Preprocessing

The application of the preprocessing strategy led to improved efficiency and quality of the data. Raw data are often marked by inaccuracy in terms of precision as well as by inherent noise, which tends to result in low-quality indicators. It was observed that cases of duplicate entries, outliers, and imbalance issues were present, which were detected by extensive analyses conducted during various phases of the studies. To rectify these issues, the z-score approach was used, in which a cutoff point was set for outliers at 0.050000, which made it easier to exclude outlier values and normalize all datasets. Next, the Yeo-Johnson transformation was applied to further fine-tune the data and make them more precise and clearer in nature. Finally, application of the SMOTE (Synthetic Minority Over-Sampling Technique) aided in the optimization of feature and data distribution across the data sets.

## 4. Methodology

The proposed model for WQI classification consists of various critical components that involve data collection, data pre-processing, construction of the WQI model, and performance evaluation of the model. Various machine learning methods were utilized to classify the WQI, and different performance evaluation methods were used to verify the performance of the model. Figure 2 summarizes and describes the proposed structure of the predictive WQI model.

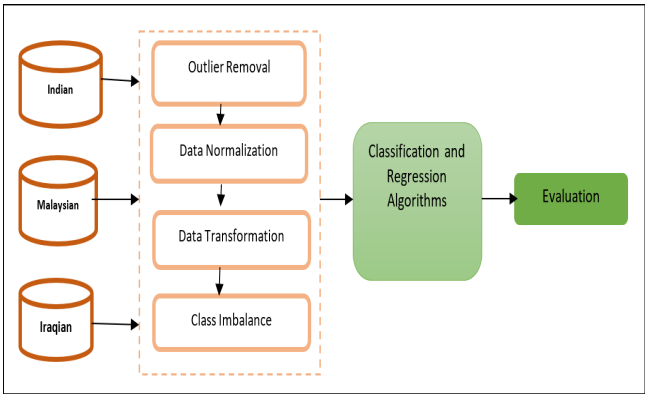


**Fig. 2:** Proposed WQI framework.

### 4.1. Building WQI Classification and Prediction Model

The method implemented in the classification of the WQI is derived from analyzing trends in a dataset related to the quality of river water. The WQI is a systematic tool for grouping data into various classes, hence providing five levels of water quality based on certain characteristics (Table 1). A diverse range of machine learning approaches, including CatBoost, SVM, Naïve Bayes, and LightGBM (LGBM), are utilized to create these classification models. Among these methodologies, CatBoost is notable owing to the use of gradient-boosted decision forests for predicting continuous variables. A relatively recent algorithm [33] is preferred because of its efficiency, accuracy, and versatility, particularly when working with categorical variables. The CatBoost algorithm works by creating an ensemble of weak decision forests, which are continually improved to enhance their predictive prowess.

In the model training stage, the metric-dependent variables

and features related to the metric were defined. Next, the core settings of essential parameters, such as the learning rate, number of trees, max-depth, and ratio of features, are defined. All of these parameters also impact the performance of the LGBM model and can be adjusted to increase performance. Model development involves the creation and training of several decision tree models, and each contributes to the improvement of the predictive accuracy of the master model.

SVM is an approach used for classification and prediction tasks. Through the conversion of every point into an n-dimensional space, SVM allows two classes to be easily linearly separable. SVMs have also been very popular in technologies, pattern recognition, and learning classification. The input space can be separated by a linear or nonlinear separating surface. In support vector classification, the separation function includes a linear combination of support vector kernels, and a decision boundary can be established.

Naïve Bayes is a classification technique grounded in Bayes' theorem., assuming that when the target variable is identified, all other variables can be assumed to be independent. This method utilizes probabilistic reasoning and statistical principles to provide predictions and classification of data. Using the combination of prior and posterior odds, the Bayesian method can avoid bias and prevent overfitting, thus increasing its ability to deal with the variability of sample data.

In forecasting water quality using the data collected in India, Malaysia, and Iraq, the following models were utilized: DTR, ELM, M5 Model, and Random Forest Regressor. They are evaluated using a complete set of performance measurements, which include RMSE, MSE, MAE, Scatter Index, Bias, and MAPE.

### 4.2. Evaluation of Performance

To find the optimal method, the performances of the models developed for WQI classification are compared. For the primary measure to choose the most efficient classification model, accuracy is used. The following statistical measures were used:

$$\text{Accuaracy } \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall } = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

In contrast, the F1-score establishes the average precision and recall values shown in Equation (8).

$$F\,1 - \mathit{Score} = \frac{2 \times \mathit{precision} \times \mathit{recall}}{\mathit{precision} + \mathit{recall}} \times 100 \tag{8}$$

In this context, TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. These metrics are significant for assessing the performance of machine learning methods in creating WQI for categorization. By using the aforementioned equations, the effectiveness of these machine learning models is examined by employing river water data to evaluate their accuracy and reliability in categorizing water quality.

Alternatively, for predictive modeling, various regression techniques are applied, such as BIAS, R2, SI, MAE, MSE, and RMSE (equation 9-14) are employed.

Bias:

$$\text{Bias}(\hat{Y}) = E(\hat{Y}) - Y \tag{9}$$

R-Squared:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \tag{10}$$

Scatter Index:

$$\text{SI} = \frac{\text{RMSE}}{\overline{X}} \tag{11}$$

Mean Absolute Error

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}| \tag{12}$$

Mean Squared Error:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y})^2 \tag{13}$$

Root Mean Squared Error:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y})^2} \tag{14}$$

## 5. Results

A study was undertaken using water quality in the Tigris-Euphrates, Klang-Langat, and Bhavani River databases to create an accurate WQI classification and build a predictive model. The databases consisted of various instances having different attribute values, and they were separated and put into training and test sets, with 80% of the instances being utilized as training and 20% as testing.

A WQI classification model was developed that included a range of machine learning techniques, including CatBoost, SVM, Naïve Bayes, and LGBM, as well as a range of regressor methods, including the ELM Regressor, M5 Model Tree, RFR, and DTR. The models were developed using a combination of independent and dependent variables. The models were tested using 20% of the dataset with the application of a range of metrics for measuring performance. In addition, the Stratified K-Fold method with 10 folds was applied on each dataset for undertaking a further detailed assessment of the performance of the models, hence ensuring a strong and statistically valid evaluation.

### 5.1. India WQI the Classification and Prediction

#### 5.1.1. Classification of WQI Models

The experimental results show that the LGBM model achieved an accuracy of 0.9772, beyond the performance of other alternative models such as Naïve Bayes, SVM, and CatBoost, whose accuracies were approximately 0.6467, 0.5734, and 0.9752, respectively. As shown in Table 2 and the supplementary materials, the LGBM model was found to perform exceptionally when compared to the other classifiers studied here. Additionally, it was observed that CatBoost, classified as a boosting algorithm, achieved high accuracy compared to the other models that had been utilized in this research.

**Table 2:** Indian Dataset Classification Accuracy

| Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC |
|---|---|---|---|---|---|---|---|
| CatBoost | 0.9752 | 0.9980 | 0.9752 | 0.9756 | 0.9752 | 0.9642 | 0.9643 |
| Naïve Bayes | 0.6467 | 0.0000 | 0.6467 | 0.6944 | 0.6596 | 0.5135 | 0.5206 |
| SVM | 0.5734 | 0.8366 | 0.5734 | 0.6568 | 0.5893 | 0.4355 | 0.4483 |
| LGBM | 0.9772 | 0.9979 | 0.9772 | 0.9776 | 0.9772 | 0.9671 | 0.9672 |

Figure 16 (Supplementary Materials) shows the ROC curve of the CatBoost algorithm applied to the Indian dataset, indicating that the results achieved by CatBoost were the second-best among the models tested. This further indicates that boosting algorithms tend to have a higher accuracy than other classification models. Figure 17 (Supplementary materials) displays the confusion matrix of the CatBoost model on the Indian dataset, where the percentage of true and predicted classes exceeded 94%. This indicates that the model performed well when using the CatBoost algorithm.

Figure 18 (Supplementary materials) shows the ROC curve of NB algorithms on Indian datasets; the results on this dataset by utilizing NB are better than those of SVM. This proves that the algorithm has a higher accuracy than the SVM classification model. However, Figure 19 (Supplementary materials) reveals the confusion matrix of NB on the Indian dataset, in which the percentage of true and predicted classes is mostly 60%, except for one, which is 39%, which means that the model performance is average.

Figure 20 (Supplementary materials) displays the confusion matrix of the SVM on the Indian dataset, where the percentage of true and predicted classes is approximately 60%. This suggests that the SVM model performed at an average level. Figure 21 (Supplementary materials) shows the ROC curve of the LGBM algorithms on the Indian datasets, where LGBM shows optimal and best results. This underscores that boosting algorithms, such as LGBM, achieve higher accuracy compared to other classification models. Additionally, Figure 22 (Supplementary Materials) presents the confusion matrix of LGBM on the Indian dataset, showing that the percentage of true and predicted classes exceeded 94%, indicating that the LGBM model performed exceptionally well.

### 5.1.2. Prediction Models

Based on the findings of the experiment, Table 3 shows that the decision tree regressor had an MAE of 1.481, the MAE for Random Forest Regressor was 1.033, the M5 Model Tree was 0.226, and the ELM Regressor was 26.47. as well as Figure 3-6.

**Table 3:** Indian Datasets Performance of prediction Model

| Model | MAE | MSE | RMSE | MAPE | R2 | S I | BIAS |
|---|---|---|---|---|---|---|---|
| DTR | 1.481 | 18.573 | 4.309 | 0.018 | 0.992 | 0.060 | -0.04 |
| RFR | 1.033 | 7.201 | 2.683 | 0.013 | 0.996 | 0.037 | -0.06 |
| M5 Model Tree | 0.226 | 0.097 | 0.312 | 0.003 | 0.999 | 0.004 | -0.00 |
| ELM Regressor | 26.475 | 1883.27 | 43.396 | 0.415 | 0.210 | 0.605 | -0.34 |

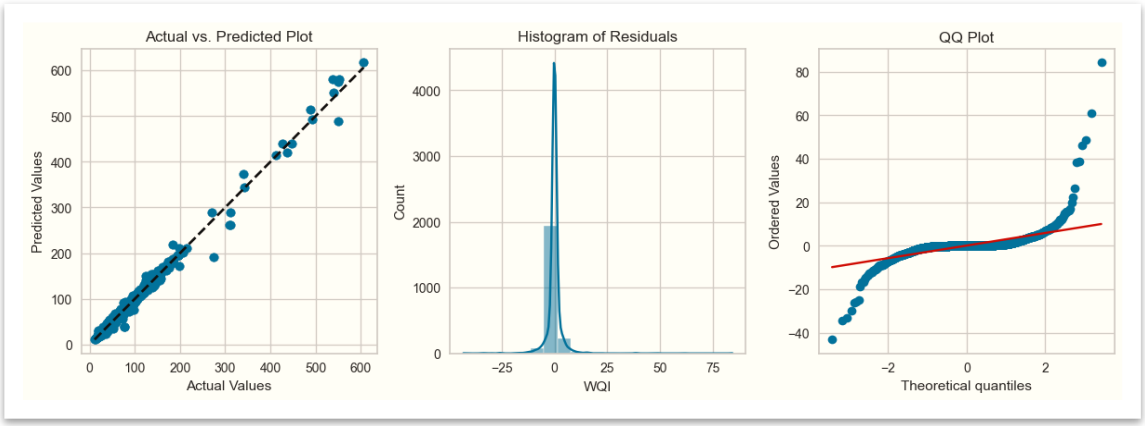


**Fig. 3:** Indian datasets plots of DTR.

Figure 3 displays the comparative performance of the M5 Model Tree on the Indian dataset, which is superior to that of both the Decision Tree Regressor and the Random Forest Regressor in all the studied parameters, which include MAE, MSE, RMSE, MAPE, Scatter Index (SI), and bias. In contrast, as can be observed in Figures 4 and 5, the Random Forest Regressor shows significantly better performance than the Decision Tree Regressor.

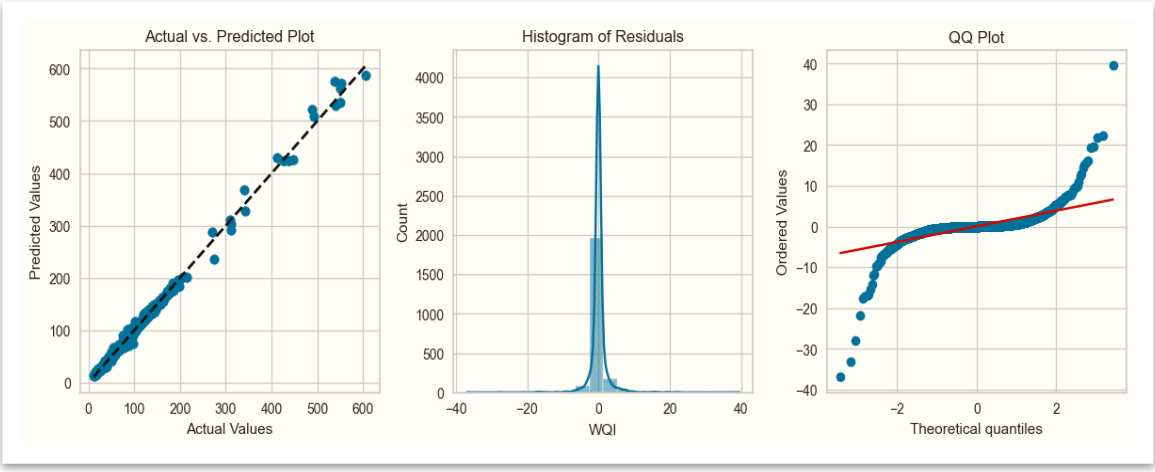


**Fig. 4:** Indian dataset plots of RFR.

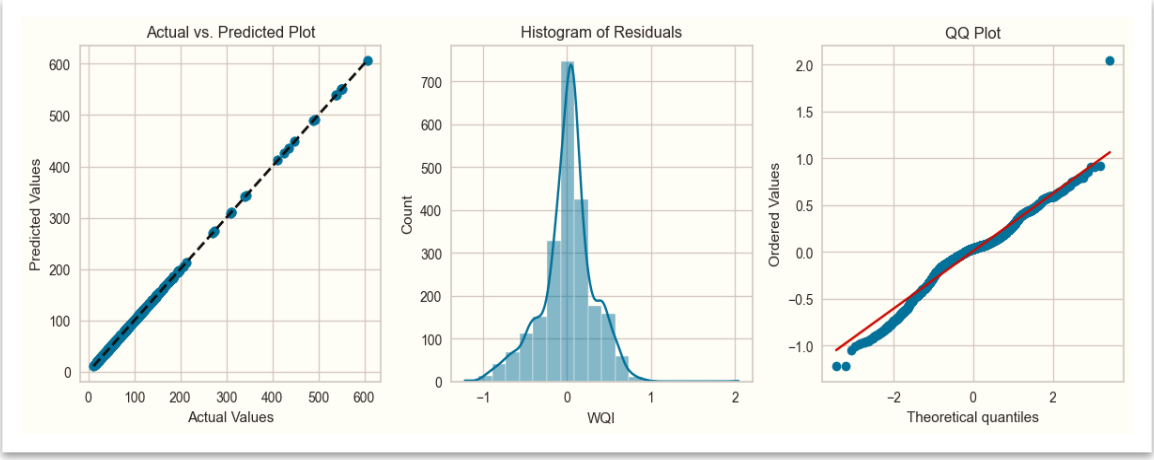


**Fig. 5:** Indian dataset plots of M5 Model Tree Regressor.

Conversely, Figure (6) illustrates the instability of the ELR on the Indian dataset.

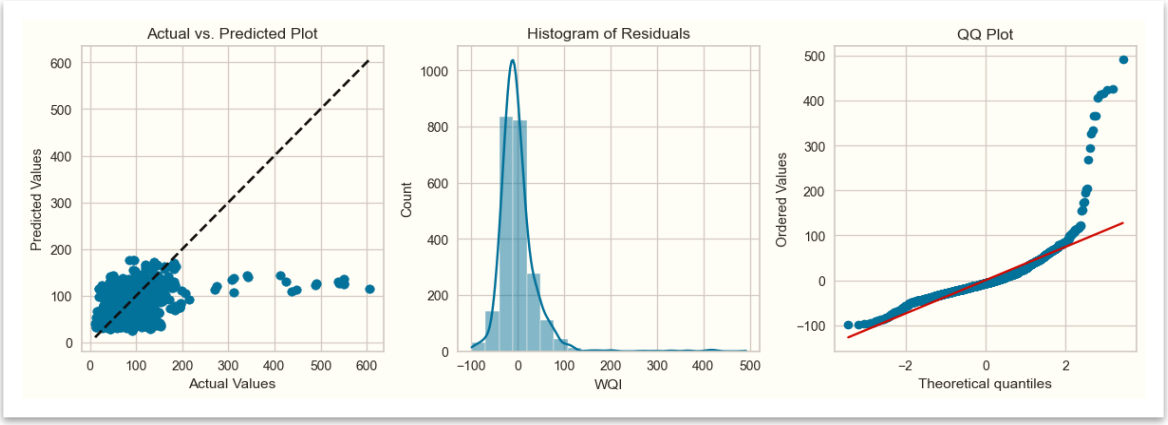


**Fig. 6:** Indian dataset plots of ELR.

### 5.2. Malaysia WQI the Classification and Prediction

This section provides an overview of the evolution of the developed classification models based on water quality data obtained from different rivers. Classifying the WQI was achieved using a combination of libraries and machine learning models, including CatBoost, SVM, Naïve Bayes, and LGBM. Under the predictive modeling approach, various sets of objectives were used to evaluate these targets using the Decision Tree Regressor, Random Forest Regressor, M5 Model Tree, and ELM Regressor. The efficacy of the WQI classification models was evaluated using various performance indicators, including accuracy, precision, recall, and F1 score. In addition, the predictive model evaluation was carried out using a wide range of indicators, including MAE, MSE, RMSE, MAPE, $R^2$, SI, and Bias.

The experimental results showed that the performance level of the LGBM and CatBoost models had accuracy levels of 0.9367 and 0.9455, respectively. In contrast, the Naïve Bayes and SVM models showed 0.7426 and 0.8885 accuracy levels, respectively. Evidence supporting the fact that CatBoost and LGBM perform better than other classifiers in terms of accuracy is given in Table 3.

Based on the experimental findings, the LGBM and CatBoost models achieved accuracies of 0.9455 and 0.9367, respectively. However, the Naïve Bayes and SVM models recorded accuracies of 0.8885 and 0.7426, respectively. As revealed in Table 3 and (Supplementary Materials), the evidence clearly shows that CatBoost and LGBM outperform other classifiers in terms of accuracy.

**Table 3:** Malaysian dataset, accuracy of classification.

| Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC |
|---|---|---|---|---|---|---|---|
| LGBM | 0.9367 | 0.9931 | 0.9367 | 0.9423 | 0.9365 | 0.9208 | 0.9222 |
| Naïve Bayes | 0.8885 | 0.9843 | 0.8885 | 0.9029 | 0.8875 | 0.8606 | 0.8645 |
| SVM | 0.7426 | 0.0000 | 0.7426 | 0.7618 | 0.7302 | 0.6781 | 0.6890 |
| CatBoost | 0.9455 | 0.9949 | 0.9455 | 0.9506 | 0.9454 | 0.9318 | 0.9331 |

Figure 23 (Supplementary materials) illustrates the ROC curve of the LGBM algorithms on Malaysian datasets; the results on this dataset using LGBM are the best results. This proves that the boosting algorithms have a higher accuracy than the other two classification models. Figure 24 (Supplementary materials) shows the confusion matrix of LGBM on the Malaysian dataset, in which the percentage of true and predicted classes is more than 95%, which means the model performs well by using this model. While Figure 25 (Supplementary materials) illustrates the ROC curve of NB algorithms on Malaysian datasets, the results on this dataset using NB were better than those using SVM. This proves that the algorithm has a higher accuracy than the SVM classification model. Figure 26 (Supplementary materials) shows the confusion matrix of NB on the Malaysian dataset, in which the percentage of true and predicted classes is more than 77%, except for one, which is 68%, indicating that the model performance is average.

Figure 27 (Supplementary materials) shows the confusion matrix of SVM on the Malaysian dataset, in which the percentage of true and predicted classes is mostly 60%, which means this model is also performing averagely.

Figure 28 (Supplementary materials) illustrates the ROC curve of the CatBoost algorithms on Malaysian datasets; the results on this dataset using CatBoost are optimal and best results. This proves that the boosting algorithms have a higher accuracy than the other two classification models. Figure 29 (Supplementary materials) shows the confusion matrix of CatBoost on the Malaysian dataset, in which the percentage of true and predicted classes is more than 92%, which means the model performs very well by using this model.

#### 5.2.1. Prediction Models

In this section, predictive models are formulated using river water quality data. Python libraries are utilized along with various regression approaches, the Decision Tree Regressor, Random Forest Regressor, M5 Model Tree, and Extreme Learning Machine (ELM Regressor), to predict the WQI. The performance of these models was evaluated through several performance criteria, including MAE, MSE, RMSE, MAPE, $R^2$, SI, and Bias, as described in Table 4 and demonstrated in the corresponding Figures 7-10.

**Table 4:** Malaysian Rivers Performance of Prediction Models.

| Model | MAE | MSE | RMSE | MAPE | R2 | S I | BIAS |
|---|---|---|---|---|---|---|---|
| DTR | 0.737 | 2.663 | 1.632 | 0.031 | 0.979 | 0.090 | 0.061 |
| RFR | 0.639 | 1.611 | 1.269 | 0.261 | 0.987 | 0.070 | -0.20 |
| M5 Model Tree | 4.627 | 4.127 | 6.424 | 2.797 | 0.999 | 3.54 | 9.200 |
| ELM Regressor | 3.943 | 2.803 | 5.294 | 2.291 | 0.999 | 2.92 | 4.086 |

Based on the experimental results, the MAE for the Decision Tree Regressor is 0.737, while the Random Forest Regressor achieves a lower MAE of 0.639. In contrast, the M5 Model Tree shows a significantly higher MAE of 4.627, and the ELM Regressor records an MAE of 3.943.

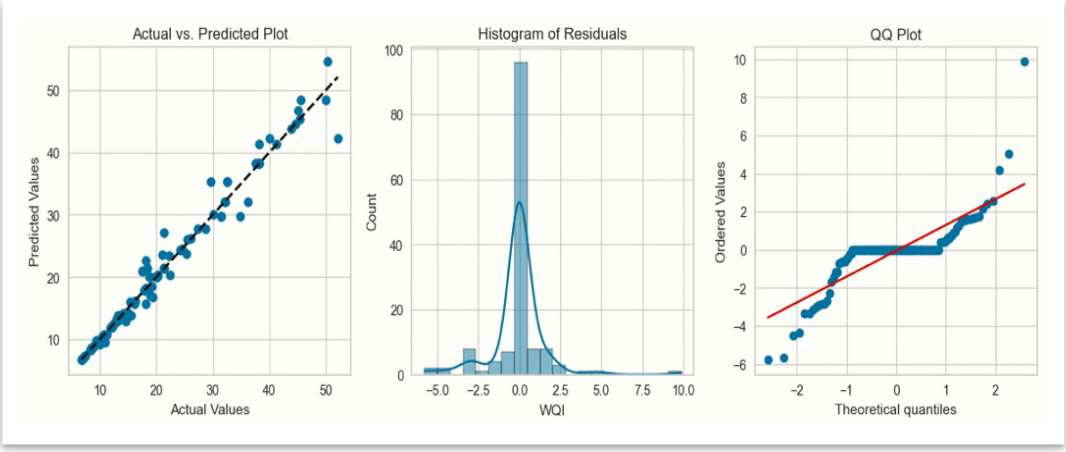


**Fig. 7:** The Malaysian dataset plots the performance of the DTR.

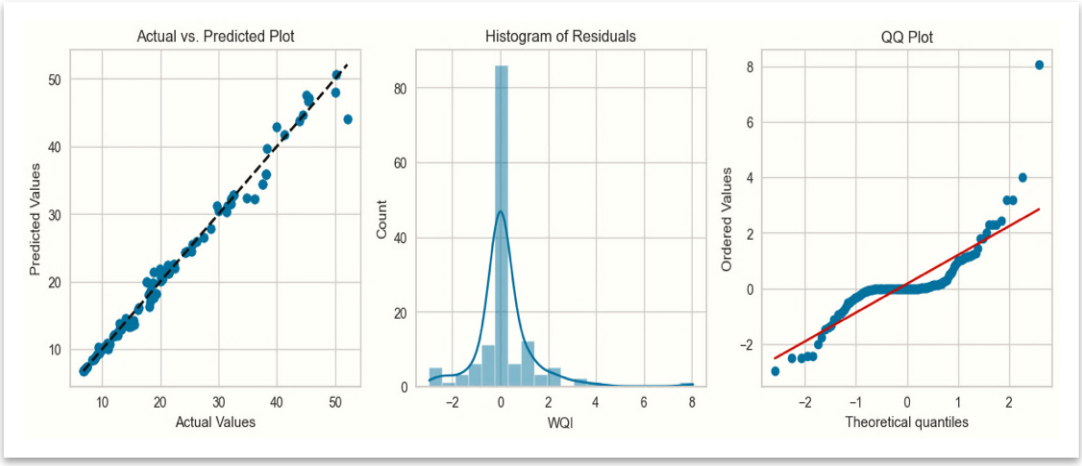


**Fig. 8:** The Malaysian dataset plots the performance of the RFR.

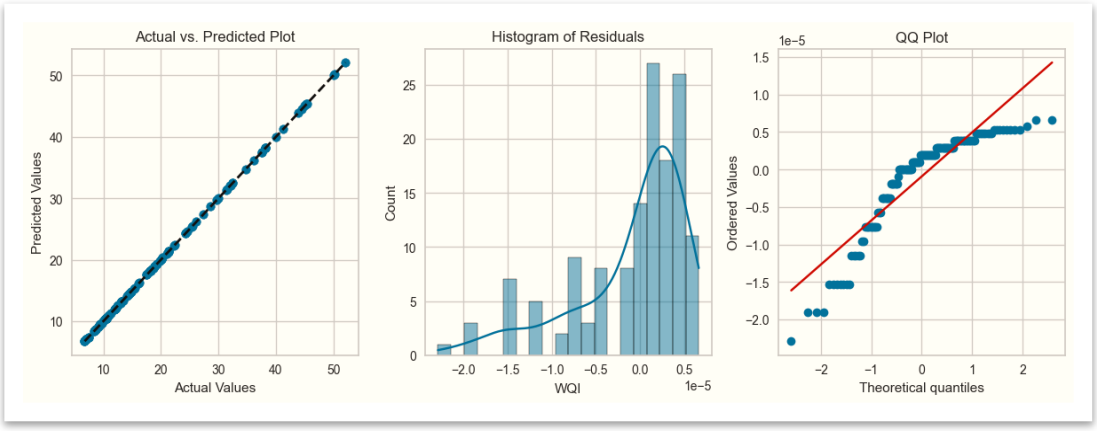


**Fig. 9:** Malaysian dataset plots of the M5 Model Tree Regressor.

As illustrated, the performance of the Decision Tree Regressor on the Malaysian dataset surpassed that of the Random Forest Regressor when evaluated using metrics such as MAPE, $R^2$, and bias. It provides more accurate and reliable predictions of water quality.

The Random Forest regressor outperformed the DTR on the Malaysian dataset according to MAE, MSE, RMSE, and SI

(Figure 9). Conversely, the Extreme Learning Machine demonstrates significantly better performance compared to the M5 Model Tree Regressor (Figure 10).

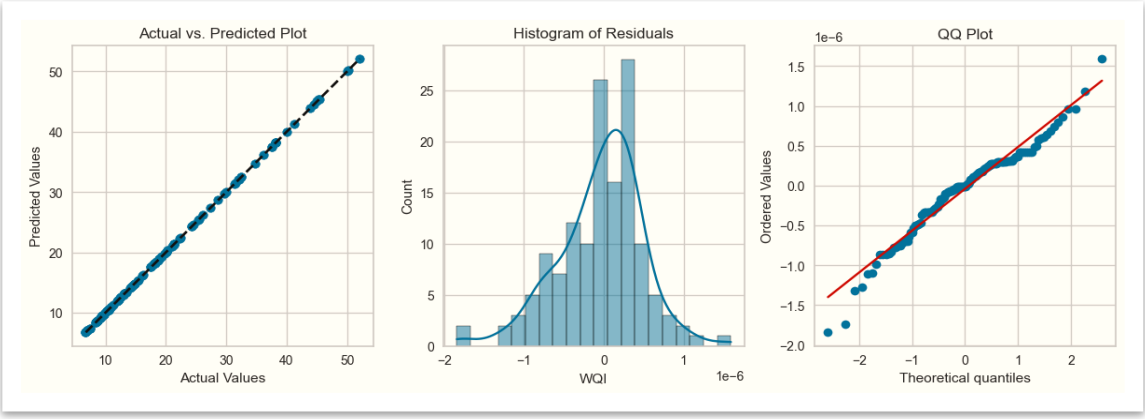


**Fig. 10:** Malaysian dataset plots of ELR.

### 5.3. Iraq WQI the Classification and Prediction

Based on the experimental results, the CatBoost model achieved an accuracy of 0.9259, whereas the accuracies for other classification models, such as Naïve Bayes, SVM, and LGBM, were 0.9110, 0.8070, and 0.9051, respectively. As shown in Table 5 and (Supplementary Materials), CatBoost demonstrates superior accuracy compared to the other classifiers. However, it is noted that CatBoost's accuracy on this dataset is lower than its performance on the Malaysian dataset, as depicted in Table 5.

**Table 5:** Iraq dataset classification accuracy.

| Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC |
|---|---|---|---|---|---|---|---|
| CatBoost | 0.9259 | 0.9698 | 0.9259 | 0.9301 | 0.9249 | 0.8877 | 0.8807 |
| Naïve Bayes | 0.9110 | 0.0000 | 0.9110 | 0.9199 | 0.9092 | 0.8647 | 0.8704 |
| SVM | 0.8070 | 0.9269 | 0.8070 | 0.8311 | 0.7917 | 0.7102 | 0.7306 |
| LGBM | 0.9051 | 0.9674 | 0.9051 | 0.9084 | 0.9043 | 0.8557 | 0.8580 |

Figure 30 (Supplementary materials) illustrates the ROC curve of CatBoost algorithms on Iraqi datasets; the results on this dataset using CatBoost are optimal and best results. This proves that the boosting algorithms have a higher accuracy than the other two classification models. Figure 31 (Supplementary materials) shows the confusion matrix of CatBoost on the Iraqi dataset, in which the percentage of true and predicted classes is more than 97%, which means the model performs very well by using this model. While Figure 32 (Supplementary materials) illustrates the ROC curve of NB algorithms on Iraqi datasets, the results on this dataset using NB are better than those using SVM. This proves that the algorithm has a higher accuracy than the SVM classification model. Whereas Figure 33 (Supplementary materials) shows the confusion matrix of NB on the Iraqi dataset, in it the percentage of true and predicted classes are mostly 84% except one which is 44% which means the model performing is averagely. Figure 34 (Supplementary materials) shows the confusion matrix of SVM on the Iraqi dataset, in which the percentage of true and predicted classes is mostly 87%, which means this model is also performing averagely. Figure 35 (Supplementary materials) illustrates the ROC curve of LGBM algorithms on Iraqi datasets; the results on this dataset using LGBM are the second-best. This proves that the boosting algorithms have a higher accuracy than the other two classification models. Figure 36 (Supplementary materials) shows the confusion matrix of LGBM on the Iraqi dataset, in which the percentage of true and predicted classes is more than 93%, which means the model performs well by using this model.

#### 5.3.1. Prediction Models

Based on the experimental results, the Decision Tree Regressor yielded an MAE of 2.66, whereas the Random Forest Regressor performed better with an MAE of 1.676. Notably, the M5 Model Tree achieved the lowest error, with an MAE of 0.002, in contrast to the ELM Regressor, which recorded a significantly higher MAE of 7.319, as displayed in Table 6 and Figures 11-14.

**Table 6:** Iraq dataset performance of the prediction model.

| Model | MAE | MSE | RMSE | MAPE | R2 | SI | BIAS |
|---|---|---|---|---|---|---|---|
| DTR | 2.66 | 19.46 | 4.412 | 1.288 | 0.909 | 0.114 | 0.745 |
| RFR | 1.676 | 10.013 | 3.164 | 1.286 | 0.953 | 0.0817 | 0.414 |
| M5 Model Tree | 0.002 | 5.669 | 0.002 | 2.306 | 0.999 | 6.154 | -3.12 |
| ELM Regressor | 7.319 | 108.484 | 10.415 | 1.950 | 0.495 | 0.269 | -2.55 |

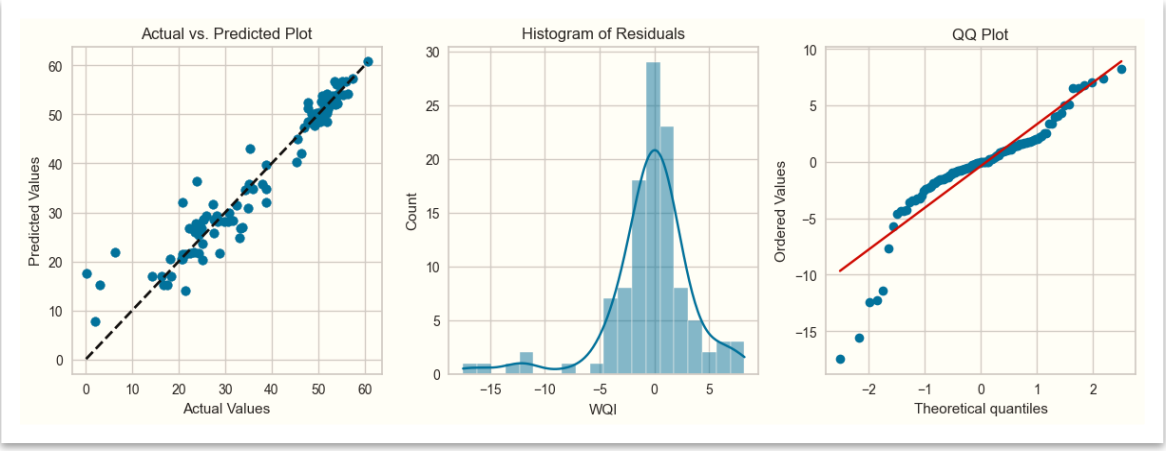


**Fig. 11:** Iraq dataset plots of the DTR.

Figure 11 shows that the DTR is quite effective when implemented on the Malaysian dataset, as it exhibits better performance when it comes to the metric $R^2$ compared to the performance of the RFR. Figure 12 shows that the RFR is more efficient than the DTR in terms of performance on the Iraqi dataset by all the performance criteria MAE, MSE, RMSE, MAPE, SI, and Bias.

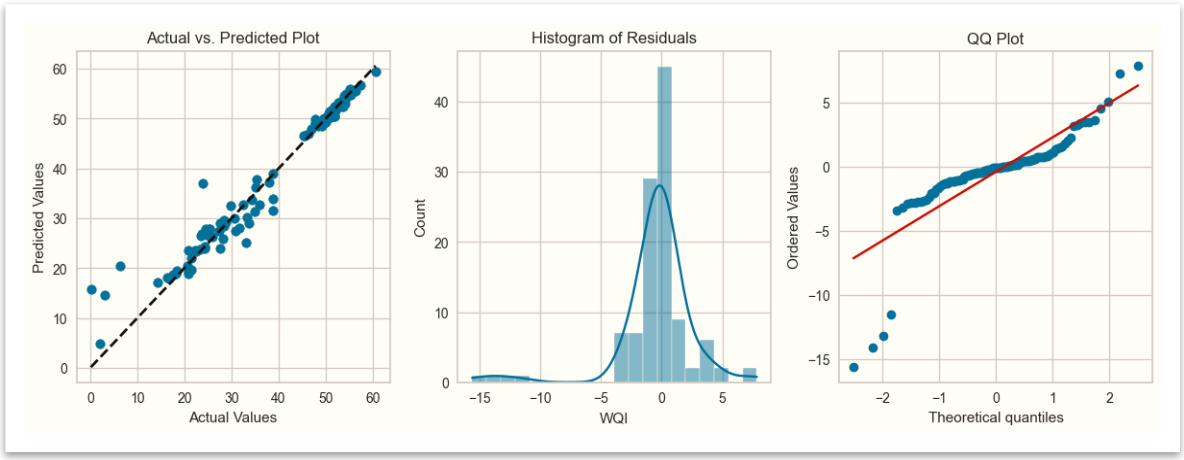


**Fig. 12:** Iraq dataset plots of RFR.

In contrast, the performance of the ELM Regressor and the M5 Model Tree when implemented on the Iraqi dataset seems to exceed expectations, providing positive performance, as shown in Figures 13 and 14.

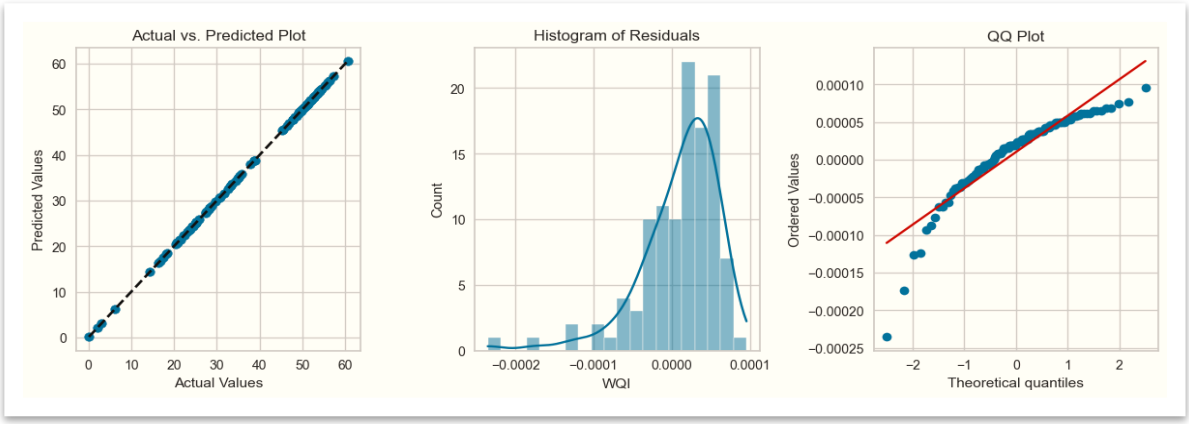


**Fig. 13:** Iraq dataset plots of the M5 Model Tree Regressor.

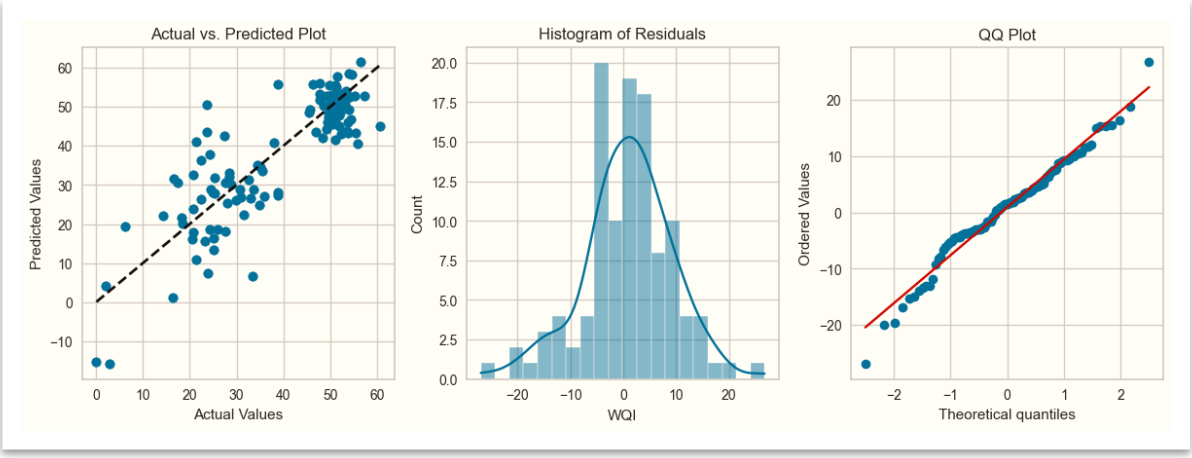


**Fig. 14:** Iraq dataset plots of ELR.

### 5.4. Discussion

The experiment involved data gathered from three distinct regions across Asia: India (South Asia), Malaysia (Southeast Asia), and Iraq (West Asia). A map generated by Google Maps indicates that these countries are predominantly covered by water, demonstrating the importance of focusing on water quality classification and prediction in these regions. Heterogeneous datasets and properties of various rivers are used to examine the effects of water quality on humans, households, agriculture, and environmental parameters. Several ML regressors and classifiers were utilized to explore the performance of the WQI. The datasets considered are Bhavani River, India; Klang and Langat, Malaysia; and Tigris and Euphrates, Iraq.

The selected datasets in three different countries introduced tremendous challenges during the machine learning model training stage, primarily because of outliers and class imbalance in the given data. To deal with these hindrances and properly prepare the dataset for successful modeling, thorough preprocessing and normalization procedures were performed. Pioneeringly, identification of outliers and correction were conducted using the Z-score approach, which set 0.050000 as a threshold for the identification and compensation of anomalous data points, thus making it possible to mold them into inliers. Next, after the elimination of the outliers, the Yeo-Johnson transformation was executed to normalize the dataset, which improved the normality of its distribution—a key requirement for the successful application of many machine learning models. Finally, to overcome the limitation of class imbalance in the data, SMOTE was used, leading to a well-balanced class distribution in the datasets. This pre-emptive measure is crucial in ensuring stable and consistent model training, as well as in increasing predictive performance and improving the machine learning models' capacity for generalization.

A number of machine learning algorithms, such as CatBoost, SVM, Naïve Bayes, and LGBM, were utilized for the classification of the WQI, where each was trained and tested thoroughly across three datasets. Algorithms such as the RF Regressor, M5 Model Tree, DT Regressor, and ELM Regressor were used for the prediction of water quality. The performance of these ML models was evaluated with a set of different measures: accuracy, AUC Curve, recall, precision, F-1 score, Kappa, and Matthew Correlation Coefficient (MCC) for classification tasks, and MAE, MSE, RMSE, MAPE, $R^2$, Scatter Index, and Bias for predictive tasks. All models showed commendable performance under a consistent experimental setup. Notably, model accuracy and MCC were primarily considered for a detailed analysis of the models' effectiveness.

The performance of the machine learning models on the Malaysian dataset was commendable, with all models achieving an accuracy of over 70%. Specifically, the SVM model recorded an accuracy of 74.26% and an MCC of 68.90%, while the Naïve Bayes model showed an improvement in accuracy by 14% but a 20% decrease in MCC compared to SVM. Notably, the CatBoost and LGBM models displayed similar performance metrics, wherein LGBM achieved an accuracy of 93.67% and an MCC of 92.22%, whereas CatBoost reached an accuracy of 94.55% and an MCC of 93.31%.

For the Indian dataset, all participating ML models performed with an accuracy exceeding 57%. The SVM model had an accuracy of 57.34% and an MCC of 44.83%, while Naïve Bayes improved upon SVM's performance with a 13% higher accuracy and an 8% higher MCC. Similarly, the CatBoost and LGBM models showed comparable results, wherein LGBM reported an accuracy of 97.72% and an MCC of 96.72%, closely followed by CatBoost with an accuracy of 97.52% and an MCC of 96.43%.

The performance of the participating ML models on the Iraq dataset was excellent with an accuracy of 80%. The accuracy and MCC of SVM model are 80.70 % and 73.06 % whereas the accuracy of Naïve Bayes is 10% better and MCC is 14% higher than SVM. On the other hand, there

are many similarities between CatBoost and LGBM; the accuracy of LGBM is 90.51% and MCC is 85.80 %, whereas the Accuracy of CatBoost is 92.59% and MCC is 88.07%.

A comparative study among four different regression models using three sets of datasets showed that every model has certain strengths depending on the corresponding features. From the results, it can be identified that the M5 Model Tree regressor is the best among all predictive models to be used on all sets of data except the Malaysian one. In this case, the Random Forest regressor shows improved performance as far as MAE, MSE, RMSE, and SI are concerned. However, the Malaysian dataset bias values are not conclusive enough to support any one of the decision trees or the Random Forest; however, the M5 Model Tree is characterized by an extremely high bias of 9.200.

Within the analysis of the Iraqi dataset, the M5 Model Tree reflects consistently improved performance measures, including MAE, MSE, RMSE, R2, and Bias. On the contrary, the Decision Tree regressor reflects superior performance in response to MAPE and SI. In the case of the Indian dataset, however, the M5 Model Tree reflects the best performance and crosses all the other regression models against the various performance measures. However, despite these strengths, the M5 Model Tree does not perform as effectively against smaller alternatives. The percentage of bias recorded for the M5 Model Tree when evaluated against the Malaysian dataset was significantly high, whereas the performance of the SI regarding the Iraqi dataset showed considerable superiority compared to other regression models. The results suggest that while the M5 Model Tree has excellent effectiveness in some situations, its dependability could be questioned in situations where model compactness and low variability of bias are of concern.

The findings list the performance of all eight ML classifiers and repressors experimented on different datasets, specifying their optimum performance and offering helpful insights on the application of machine learning on datasets for water quality evaluation. The employment of metrics such as SI and Bias in the analysis gives researchers a more insightful understanding of the behavior of machine learning models with data and the subsequent effect on results. The evidence suggests variation in the WQI among different countries, suggesting that the Malaysian-Iraqi water quality is better overall, which will be advantageous to households, industries, agriculture, and the environment. There are fewer environmental issues resulting from better water quality, which is suitable for nearby countries and the environment in general.

Our results show that the selected algorithms of CatBoost, LightGBM, SVM, Naïve Bayes, Random Forest, M5P, and DT align closely with the top-performing models reported in WQI studies. For classification, the ensemble and boosting methods consistently ranked among the highest. For example, Singh *et al.* (2025) [40] (Telangana, India) found that a soft-voting ensemble (combining DT, LR, and SVM) achieved 96.39% accuracy (F1 = 96.41%). In that study, SVM alone achieved approximately 95.7% accuracy and DT 95.0%. Likewise, analysis of the Langat River (Malaysia) dataset showed that SVM produced the best results, with 96.35% macro accuracy and ~82.4% F1, compared to 94.71% accuracy and ~81.4% F1 for DT [41].

Gradient-boosting algorithms similarly demonstrate strong performance: Karthick *et al.* (2024) reported XGBoost ≈96.31% accuracy and noted that SMOTE further enhanced CatBoost's results [42]. Torky *et al.* (2023) likewise found that LGBM achieved 97% test accuracy on ~7996 samples on data from Johor, Malaysia [43].

Boosting methods also lead to regression or continuous WQI prediction. Choudhary *et al.* (2025) (India) achieved $R^2 \approx 0.9952$ using a stacked ensemble and $R^2 \approx 0.9894$ by means of CatBoost (RMSE ≈ 1.59) [44]. Other studies similarly reported high $R^2$ and low error for tree-based models. For example, Islam *et al.* (2024) (Bangladesh) found Gradient Boosting $R^2 \approx 0.97$ and Random Forest $R^2 \approx 0.96$ [45]. Simpler models, such as Naïve Bayes, rarely reach these benchmarks and were not highlighted in most studies. Importantly, research reporting MCC shows comparably high values. For example, Singh *et al.* (2025) recorded an ensemble MCC ≈ 0.9308 with an accuracy of 96.39% [40].

The outcomes illuminate the performance of all eight machine learning classifiers and all eight methods of regression in different datasets, in turn defining their best performance and providing insights on how machine learning can be used in determining water quality. The use of indicators such as SI and Bias provides the study with better clarity by enabling a better understanding of the performance of machine learning models on the data, as well as the implications of the results. The heterogeneity of WQI in different countries suggests that water quality in Malaysia is relatively high compared to that of Iraq, hence providing benefits to households, industries, agriculture, as well as the environment. High environmental quality is accompanied by a decline in environmental problems, thereby benefitting neighboring countries as well as the overall environment worldwide.

## 6. Conclusion

The present study was conducted to examine the WQI within Southeast Asian, South Asian, and West Asian areas and its influence on the population, households, and environmental conditions, including surrounding areas. Multitude of machine learning algorithms were utilized during the classification and forecasting of the WQI, including CatBoost, SVM, Naïve Bayes, and LightGBM as classification models; and RFR, M5 Model Tree, DTR, and EML Regressor as forecasting models, using river data

from Malaysia, India, and Iraq. Critical variables, including temperature, nitrate concentration, pH, BOD, DO, and TC, were examined throughout the study. The accuracy of these models when classifying and forecasting river WQI was compared using various performance indicators. Of notable interest, CatBoost and LightGBM displayed high accuracy when classifying various sets of data, significantly outperforming the other models when classifying the water quality index. Future research should involve blending other boost family algorithms and ensemble approaches to enhance the accuracy of classifying and forecasting the quality of water.

**References**


[1] D. Hinrichsen and H. Tacio, "The Coming Freshwater Crisis is Already Here."

[2] A. Litke and A. Rieu-Clarke, "The UN Watercourses Convention and its complementary User's Guide-Indispensable ingredients for global water cooperation," in *Global Water: Issues and Insights*, Dec. 16, 2023., vol. Accessed: ANU Press, 2023. doi: 10.22459/GW.05.2014.36.

[3] R. Das Kangabam and M. Govindaraju, "Anthropogenic activity-induced water quality degradation in the Loktak lake, a Ramsar site in the Indo-Burma biodiversity hotspot," *Environ Technol*, vol. 40, no. 17, pp. 2232–2241, Jul. 2019, doi: 10.1080/09593330.2017.1378267.

[4] X. Nong, D. Shao, H. Zhong, and J. Liang, "Evaluation of water quality in the South-to-North Water Diversion Project of China using the water quality index (WQI) method," *Water Res*. 178, p. 115781, Jul. 2020, doi: 10.1016/j.watres.2020.115781.

[5] M. Anul Haq, M. Abdul Rahim Khan, and M. Alshehri, "Insider Threat Detection Based on NLP Word Embedding and Machine Learning," *Intelligent Automation & Soft Computing*, vol. 33, no. 1, pp. 619–635, 2022, doi: 10.32604/iasc.2022.021430.

[6] M. Anul Haq, "Planetscope Nanosatellites Image Classification Using Machine Learning," *Computer Systems Science and Engineering*, vol. 42, no. 3, pp. 1031–1046, 2022, doi: 10.32604/csse.2022.023221.

[7] M. Anul Haq, "CNN Based Automated Weed Detection System Using UAV Imagery," *Computer Systems Science and Engineering*, vol. 42, no. 2, pp. 837–849, 2022, doi: 10.32604/csse.2022.023016.

[8] A. Attaallah and R. Ahmad Khan, "SMOTEDNN: A Novel Model for Air Pollution Forecasting and AQI Classification," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1403–1425, 2022, doi: 10.32604/cmc.2022.021968.

[9] G. Revathy, S. A. Alghamdi, S. M. Alahmari, S. R. Yonbawi, A. Kumar, and M. Anul Haq, "Sentiment analysis using machine learning: Progress in the machine intelligence for data science," *Sustainable Energy Technologies and Assessments*, vol. 53, p. 102557, Oct. 2022, doi: 10.1016/j.seta.2022.102557.

[10] B. P. Santosh Kumar *et al.*, "Fine-tuned convolutional neural network for different cardiac view classification," *J Supercomput*, vol. 78, no. 16, pp. 18318–18335, Nov. 2022, doi: 10.1007/s11227-022-04587-0.

[11] S. Khullar and N. Singh, "Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation," *Environmental Science and Pollution Research*, vol. 29, no. 9, pp. 12875–12889, Feb. 2022, doi: 10.1007/s11356-021-13875-w.

[12] V. V. P. D, L. Y Venkataramana, P. S. Kumar, P. G, S. K., and P. A.J., "Water quality analysis in a lake using deep learning methodology: prediction and validation," *Int J Environ Anal Chem*, vol. 102, no. 17, pp. 5641–5656, Dec. 2022, doi: 10.1080/03067319.2020.1801665.

[13] H. Shaheed, M. H. Zawawi, and G. Hayder, "Water Quality Index Classification of Southeast, South and West Asia Rivers using Machine Learning Algorithms," *Journal of Ecohumanism*, vol. 3, no. 8, Nov. 2024, doi: 10.62754/joe.v3i8.4750.

[14] J. P. Nair and M. S. Vijaya, "River Water Quality Prediction and index classification using Machine Learning," *J Phys Conf Ser*, vol. 2325, no. 1, p. 012011, Aug. 2022, doi: 10.1088/1742-6596/2325/1/012011.

[15] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: an experimental review," *J Big Data*, vol. 7, no. 1, p. 70, Dec. 2020, doi: 10.1186/s40537-020-00349-y.

[16] C. Sillberg, P. Kullavanijaya, and O. Chavalparit, "Water Quality Classification by Integration of Attribute-Realization and Support Vector Machine for the Chao Phraya River," *Journal of Ecological Engineering*, vol. 22, no. 9, pp. 70–86, Oct. 2021, doi: 10.12911/22998993/141364.

[17] M. Yilma, Z. Kiflie, A. Windsperger, and N. Gessese, "Application of artificial neural network in water quality index prediction: a case study in Little Akaki River, Addis Ababa, Ethiopia," *Model Earth Syst Environ*, vol. 4, no. 1, pp. 175–187, Apr. 2018, doi: 10.1007/s40808-018-0437-x.

[18] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, "Efficient Water Quality Prediction Using Supervised Machine Learning," *Water (Basel)*, vol. 11, no. 11, p. 2210, Oct. 2019, doi: 10.3390/w11112210.

[19] M. Sakizadeh, "Artificial intelligence for the

prediction of water quality index in groundwater systems," *Model Earth Syst Environ*, vol. 2, no. 1, p. 8, Mar. 2016, doi: 10.1007/s40808-015-0063-9.

[20] J. P. Nair and M. S. Vijaya, "Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A Survey," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, IEEE, Mar. 2021, pp. 1747–1753. doi: 10.1109/ICAIS50930.2021.9395832.

[21] H. Shaheed, M. H. Zawawi, and G. Hayder, "The Development of a River Quality Prediction Model That Is Based on the Water Quality Index via Machine Learning: A Review," *Processes*, vol. 13, no. 3, p. 810, Mar. 2025, doi: 10.3390/pr13030810.

[22] J. P. Nair and V. M. S Associate Professor, "Analysing And Modelling Dissolved Oxygen Concentration Using Deep Learning Architectures," *International Journal of Mechanical Engineering*, vol. 7, no. 07, 2022, doi: 10.56452/2022-11-002.

[23] J. P. Nair and M. S. Vijaya, "DESIGN AND DEVELOPMENT OF EFFICIENT WATER QUALITY PREDICTION MODELS USING VARIANTS OF RECURRENT NEURAL NETWORKS," *europeanchemicalbulletin*, vol. 12, no. si5, Oct. 2023, doi: 10.48047/ecb/2023.12.si5.0143.

[24] J. P. Nair and V. M. S, "Temporal Fusion Transformer: A Deep Learning Approach for Modeling and Forecasting River Water Quality Index International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Temporal Fusion Transformer: A Deep Learning Approach for Modeling and Forecasting River Water Quality Index." [Online]. Available: https://www.researchgate.net/publication/373216517

[25] N. H. A. Malek, W. F. Wan Yaacob, S. A. Md Nasir, and N. Shaadan, "Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques," *Water (Basel)*, vol. 14, no. 7, p. 1067, Mar. 2022, doi: 10.3390/w14071067.

[26] A. Fernández del Castillo, C. Yebra-Montes, M. Verduzco Garibay, J. de Anda, A. Garcia-Gonzalez, and M. S. Gradilla-Hernández, "Simple Prediction of an Ecosystem-Specific Water Quality Index and the Water Quality Classification of a Highly Polluted River through Supervised Machine Learning," *Water (Basel)*, vol. 14, no. 8, p. 1235, Apr. 2022, doi: 10.3390/w14081235.

[27] J. Xia and J. Zeng, "Environmental factor assisted chlorophyll-a prediction and water quality eutrophication grade classification: a comparative analysis of multiple hybrid models based on a SVM," *Environ Sci (Camb)*, vol. 7, no. 6, pp. 1040–1049, 2021, doi: 10.1039/D0EW01110J.

[28] L. Sheng, J. Zhou, X. Li, Y. Pan, and L. Liu, "Water quality prediction method based on preferred classification," *IET Cyber-Physical Systems: Theory & Applications*, vol. 5, no. 2, pp. 176–180, Jun. 2020, doi: 10.1049/iet-cps.2019.0062.

[29] D. H. Nguyen, X. Hien Le, J.-Y. Heo, and D.-H. Bae, "Development of an Extreme Gradient Boosting Model Integrated With Evolutionary Algorithms for Hourly Water Level Prediction," *IEEE Access*, vol. 9, pp. 125853–125867, 2021, doi: 10.1109/ACCESS.2021.3111287.

[30] A. D. Martinho, H. S. Hippert, and L. Goliatt, "Short-term streamflow modeling using data-intelligence evolutionary machine learning models," *Sci Rep*, vol. 13, no. 1, p. 13824, Aug. 2023, doi: 10.1038/s41598-023-41113-5.

[31] M. N. Adli Zakaria *et al.*, "Exploring machine learning algorithms for accurate water level forecasting in Muda river, Malaysia," *Heliyon*, vol. 9, no. 7, p. e17689, Jul. 2023, doi: 10.1016/j.heliyon.2023.e17689.

[32] H. A. Karim Al-Jaf, "Water Quality Index Application to Evaluate the Ground Water Quality in Kalar City- Kurdistan Region- Iraq," *IOP Conf Ser Earth Environ Sci*, vol. 1120, no. 1, p. 012002, Dec. 2022, doi: 10.1088/1755-1315/1120/1/012002.

[33] V. Kumar, N. Kedam, K. V. Sharma, D. J. Mehta, and T. Caloiero, "Advanced Machine Learning Techniques to Improve Hydrological Prediction: A Comparative Analysis of Streamflow Prediction Models," *Water (Basel)*, vol. 15, no. 14, p. 2572, Jul. 2023, doi: 10.3390/w15142572.

[34] W. Yee Wong *et al.*, "A Stacked Ensemble Deep Learning Approach for Imbalanced Multi-Class Water Quality Index Prediction," *Computers, Materials & Continua*, vol. 76, no. 2, pp. 1361–1384, 2023, doi: 10.32604/cmc.2023.038045.

[35] A. L. Lopez, N. A. Haripriya, K. Raveendran, S. Baby, and C. V Priya, "Water quality prediction system using LSTM NN and IoT," in *2021 IEEE International Power and Renewable Energy Conference (IPRECON)*, IEEE, Sep. 2021, pp. 1–6. doi: 10.1109/IPRECON52453.2021.9640938.

[36] M. A. Rahu, A. F. Chandio, K. Aurangzeb, S. Karim, M. Alhussein, and M. S. Anwar, "Toward Design of Internet of Things and Machine Learning-Enabled Frameworks for Analysis and Prediction of Water Quality," *IEEE Access*, vol. 11, pp. 101055–101086, 2023, doi: 10.1109/ACCESS.2023.3315649.

[37] M. Ahmed, R. Mumtaz, Z. Anwar, and S. M. H.

Zaidi, "Assessment of the monsoonal impact of air pollutants and meteorological factors on physicochemical water quality parameters using remote sensing," *Journal of Water and Climate Change*, vol. 14, no. 7, pp. 2164–2190, Jul. 2023, doi: 10.2166/wcc.2023.500.

[38] M. Ahmed, R. Mumtaz, and Z. Anwar, "An Enhanced Water Quality Index for Water Quality Monitoring Using Remote Sensing and Machine Learning," *Applied Sciences*, vol. 12, no. 24, p. 12787, Dec. 2022, doi: 10.3390/app122412787.

[39] M. YURTSEVER and M. EMEÇ, "Potable Water Quality Prediction Using Artificial Intelligence and Machine Learning Algorithms for Better Sustainability," *Ege Akademik Bakis (Ege Academic Review)*, Mar. 2023, doi: 10.21121/eab.1252167.

[40] P. Singh *et al.*, "An ensemble-driven machine learning framework for enhanced water quality classification," *Discover Sustainability*, vol. 6, no. 1, p. 552, Jun. 2025, doi: 10.1007/s43621-025-01467-4.

[41] I. I. S. Shamsuddin, Z. Othman, and N. S. Sani, "Water Quality Index Classification Based on Machine Learning: A Case from the Langat River Basin Model," *Water (Basel)*, vol. 14, no. 19, p. 2939, Sep. 2022, doi: 10.3390/w14192939.

[42] K. K, S. Krishnan, and R. Manikandan, "Water quality prediction: a data-driven approach exploiting advanced machine learning algorithms with data augmentation," *Journal of Water and Climate Change*, vol. 15, no. 2, pp. 431–452, Feb. 2024, doi: 10.2166/wcc.2023.403.

[43] M. Torky, A. Bakhiet, M. Bakrey, A. A. Ismail, and A. I. B. E. L. Seddawy, "Recognizing safe drinking water and predicting water quality index using machine learning framework," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 1, 2023.

[44] R. Choudhary, A. Kumar, P. C., M. M. Naik, M. Choudhury, and N. A. Khan, "Predicting water quality index using stacked ensemble regression and SHAP based explainable artificial intelligence," *Sci Rep*, vol. 15, no. 1, p. 31139, Aug. 2025, doi: 10.1038/s41598-025-09463-4.

[45] M. J. Islam *et al.*, "Machine Learning-Driven Water Quality Index Prediction: Enhancing Accuracy with Gradient Boosting and Explainable AI for Sustainable Water Monitoring," *Applied Agriculture Sciences*, vol. 2, no. 1, pp. 1–14, 2024.

[46] Latif, M., Nasir, N., Nawaz, R. et al. Assessment of drinking water quality using Water Quality Index and synthetic pollution index in urban areas of mega city Lahore: a GIS-based approach. Sci Rep 14, 13416 (2024). https://doi.org/10.1038/s41598-024-63296-1

**Supplementary materials:**

**Bhavani River of India**

**Table 1:** Bhavani River India Dataset Statistics

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | Temp | 7648 non-null | float64 |
| 1 | pH | 7648 non-null | float64 |
| 2 | Conductivity | 7648 non-null | float64 |
| 3 | Turbidity | 7648 non-null | float64 |
| 4 | PhenolphthAlkalinity | 7648 non-null | float64 |
| 5 | Total Alkalinity | 7648 non-null | float64 |
| 6 | Cholride | 7648 non-null | float64 |
| 7 | COD | 7648 non-null | float64 |
| 8 | TKN | 7648 non-null | float64 |
| 9 | Ammonia | 7648 non-null | float64 |
| 10 | Hardness | 7648 non-null | float64 |
| 11 | Ca.Hardness | 7648 non-null | float64 |
| 12 | Mg.Hardness | 7648 non-null | float64 |
| 13 | Sulphate | 7648 non-null | float64 |
| 14 | Sodium | 7648 non-null | float64 |
| 15 | TSS | 7648 non-null | int64 |
| 16 | TDS | 7648 non-null | float64 |
| 17 | FDS | 7648 non-null | float64 |
| 18 | Phosphate | 7648 non-null | float64 |
| 19 | Boron | 7648 non-null | float64 |
| 20 | Pottassium | 7648 non-null | float64 |
| 21 | BOD | 7648 non-null | float64 |
| 22 | Fluoride | 7648 non-null | float64 |
| 23 | Nitrate-N | 7648 non-null | float64 |
| 24 | TC | 7648 non-null | float64 |
| 25 | FC | 7648 non-null | float64 |
| 26 | Dew | 7646 non-null | float64 |
| 27 | Humidity | 7646 non-null | float64 |
| 28 | Sealevelpressure | 7646 non-null | float64 |
| 29 | Precipitation | 7646 non-null | float64 |
| 30 | Precipcover | 7646 non-null | float64 |
| 31 | Windspeed | 7646 non-null | float64 |
| 32 | Winddir | 7646 non-null | float64 |
| 33 | Cloudcover | 7646 non-null | float64 |
| 34 | Visibility | 7646 non-null | float64 |
| 35 | Station | 7648 non-null | int64 |
| 36 | Latitude | 7648 non-null | object |
| 37 | Longitude | 7648 non-null | object |
| 38 | Year | 7648 non-null | int64 |
| 39 | Date | 7648 non-null | object |
| 40 | DO | 7648 non-null | float64 |
| 41 | WQI | 7647 non-null | float64 |
| 42 | WQC | 7647 non-null | float64 |

**Klang-Langar Rivers of Malaysia**

**Table 2:** Dataset Statistics

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | DO | 655 non-null | float64 |
| 1 | BOD | 655 non-null | int64 |
| 2 | COD | 655 non-null | int64 |
| 3 | SS | 655 non-null | int64 |
| 4 | pH | 655 non-null | float64 |
| 5 | NH3-N | 655 non-null | float64 |
| 6 | WQC1 | 655 non-null | int64 |
| 7 | WQI | 655 non-null | float64 |
| 8 | WQC | 655 non-null | object |

dtypes: float64(4), int64(4), object(1)

**Iraqi Tigris and Euphrates Rivers**

**Table 3:** Dataset Statistics

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | Q m3/s | 376 non-null | int64 |
| 1 | PH | 376 non-null | float64 |
| 2 | Temp | 376 non-null | float64 |
| 3 | DO2 | 376 non-null | float64 |
| 4 | PO4 | 376 non-null | float64 |
| 5 | NO3 | 376 non-null | float64 |
| 6 | Ca | 376 non-null | float64 |
| 7 | Mg | 376 non-null | float64 |
| 8 | TH | 376 non-null | float64 |
| 9 | K | 376 non-null | float64 |
| 10 | Na | 376 non-null | float64 |
| 11 | SO4 | 376 non-null | float64 |
| 12 | CL | 376 non-null | float64 |
| 13 | TDS | 376 non-null | float64 |
| 14 | EC | 376 non-null | float64 |
| 15 | Alk | 376 non-null | float64 |
| 16 | WQI | 376 non-null | float64 |

**Figure 1**: Map of the Bhavani River in India.

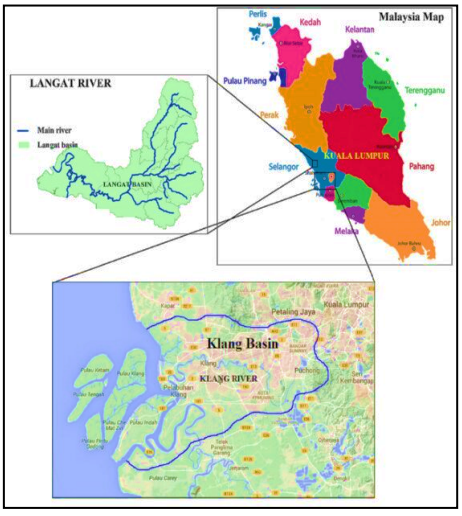


**Figure 2:** Map of Klang and Langat Rivers in Malaysia

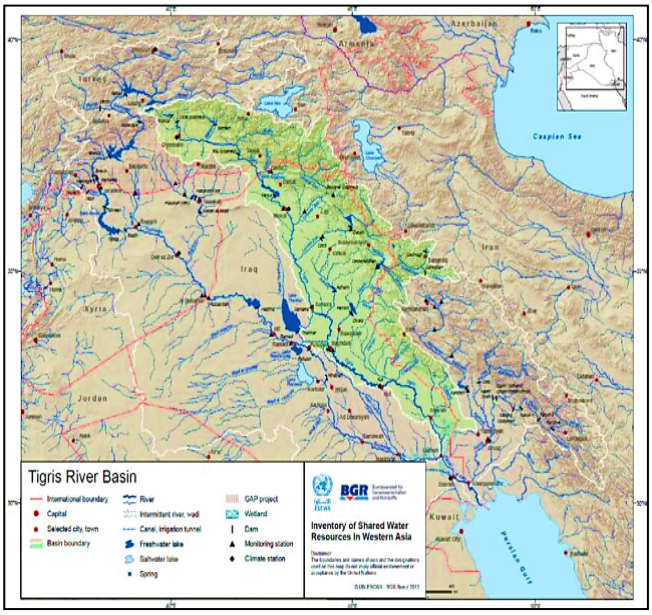


**Figure 3:** Map of the Tigris and Euphrates Rivers in Iraq.

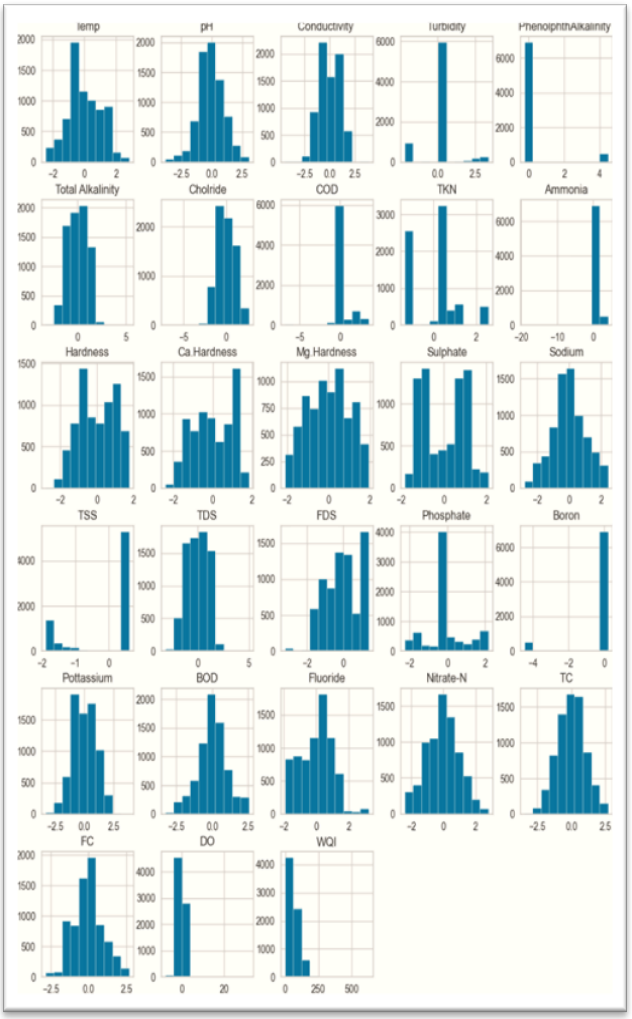


**Figure 4:** India Histogram of Bhavani River.

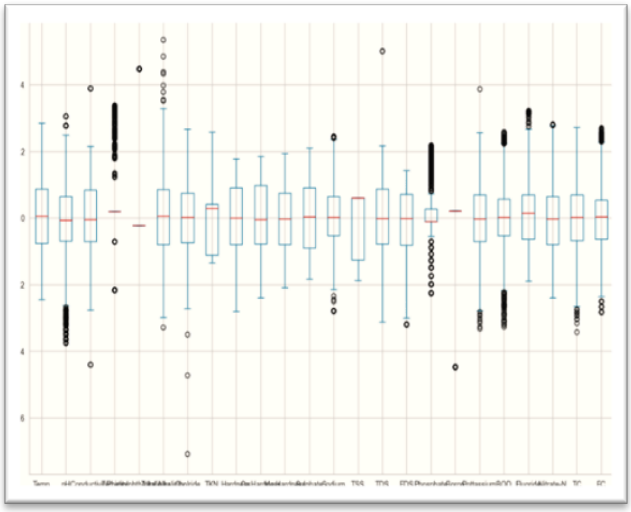

**Figure 5:** India Box Plot of Bhavani River.

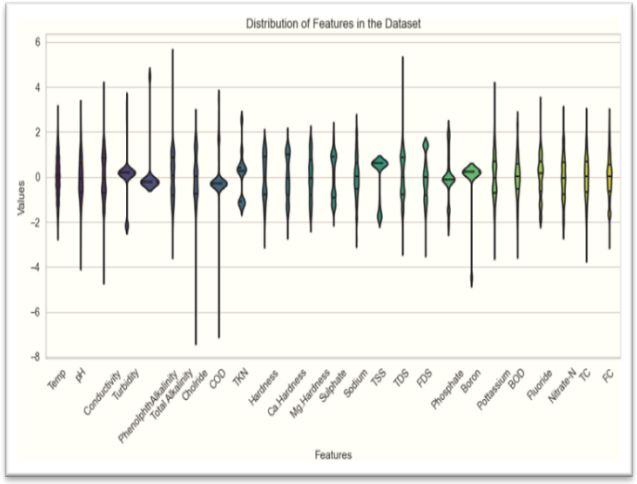


**Figure 6:** India Violin Plot of Bhavani River

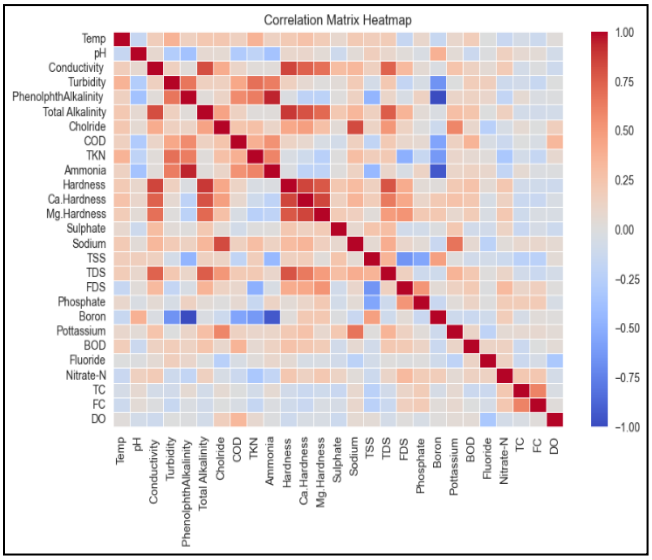


**Figure 7:** India Correlation Matrix of Bhavani River attributes.

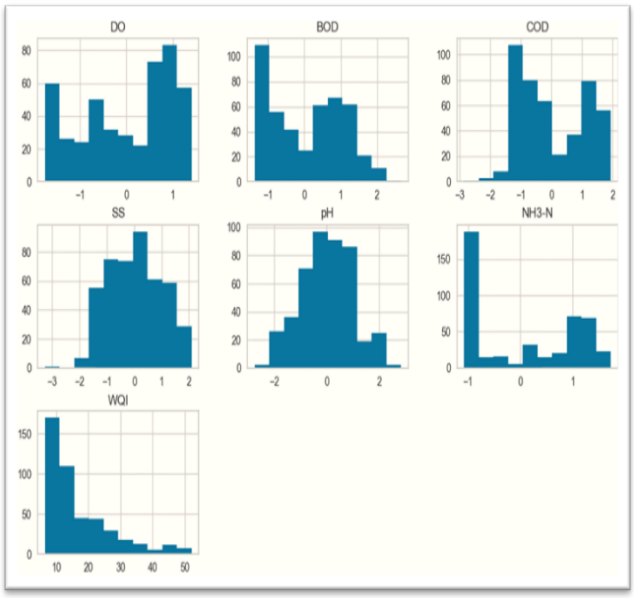


**Figure 8:** Malaysia Histogram of Klang-Langar Rivers.

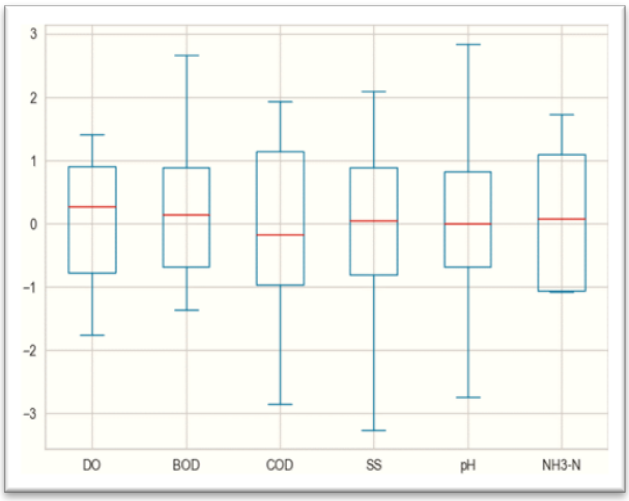


**Figure 9:** Malaysia Box Plot of Klang-Langar Rivers

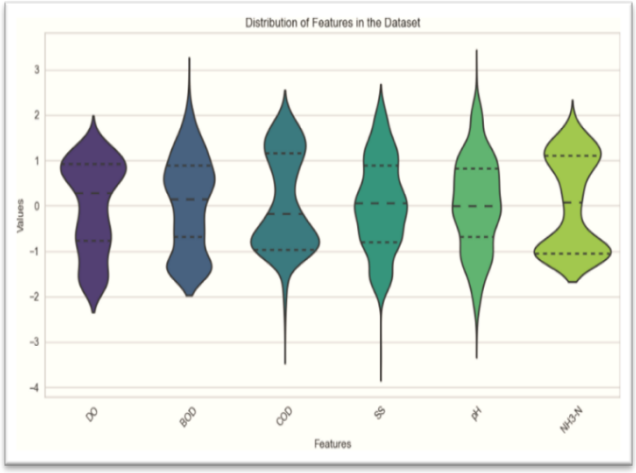


**Figure 10:** Malaysia Violin Plot of Klang-Langar Rivers.

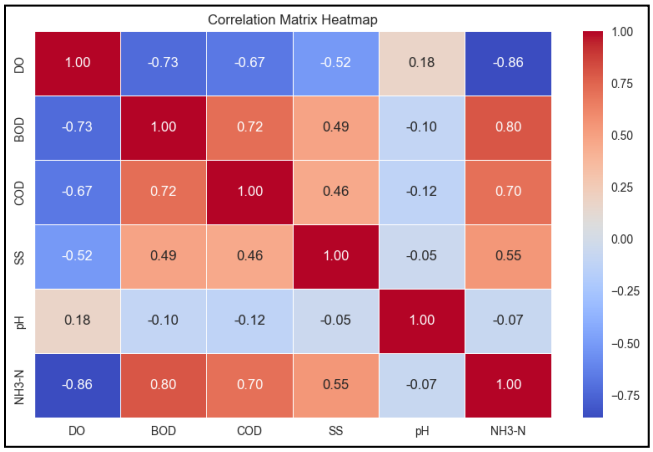


**Figure 11:** Malaysia Correlation Matrix of Klang-Langar River Parameters.

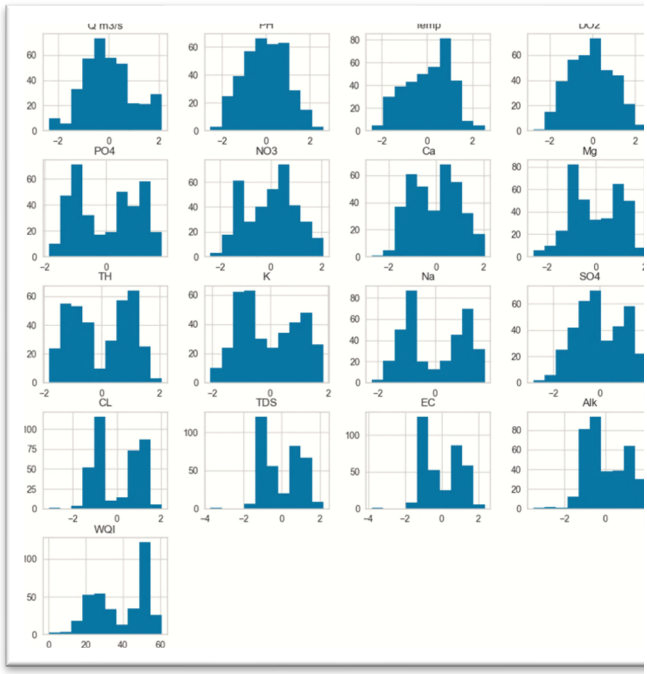


**Figure 12:** Iraq Histogram of Tigris and Euphrates Rivers.

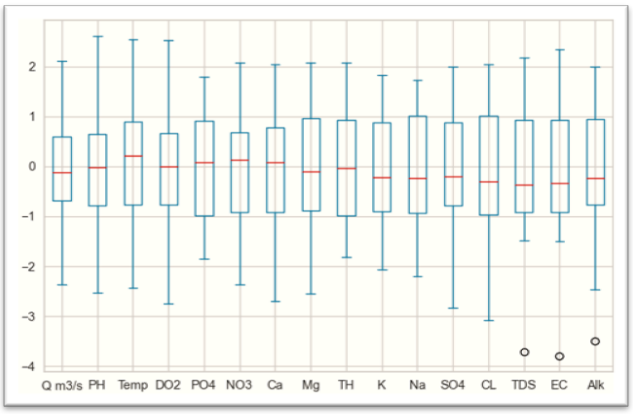


**Figure 13:** Iraq Box Plot of Tigris and Euphrates Rivers.

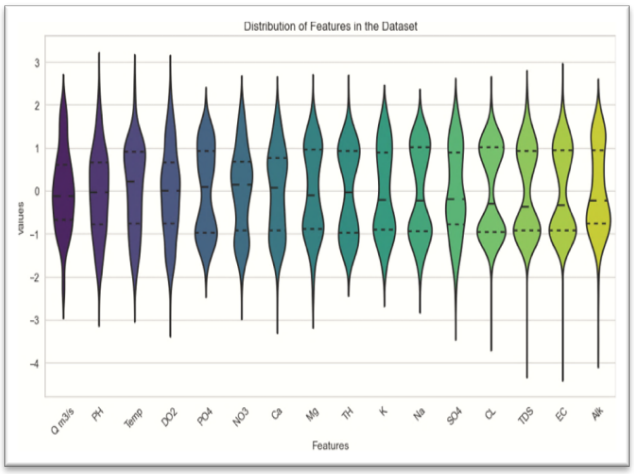


**Figure 14:** Iraq Violin Plot of Tigris and Euphrates Rivers

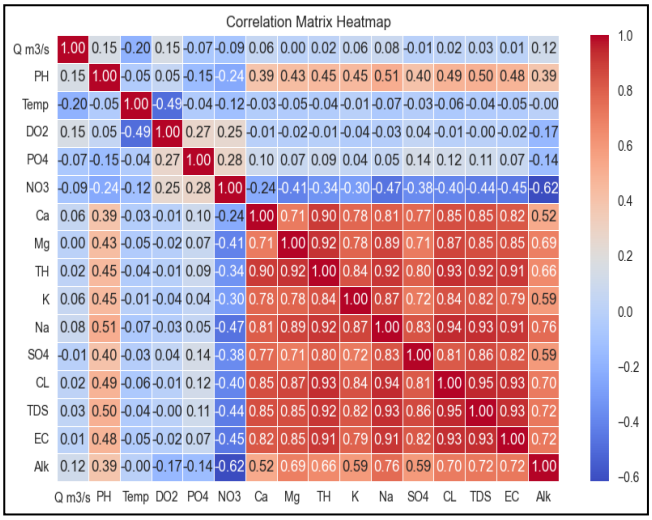


**Figure 15:** Iraq Correlation Matrix of Tigris and Euphrates River Parameters.

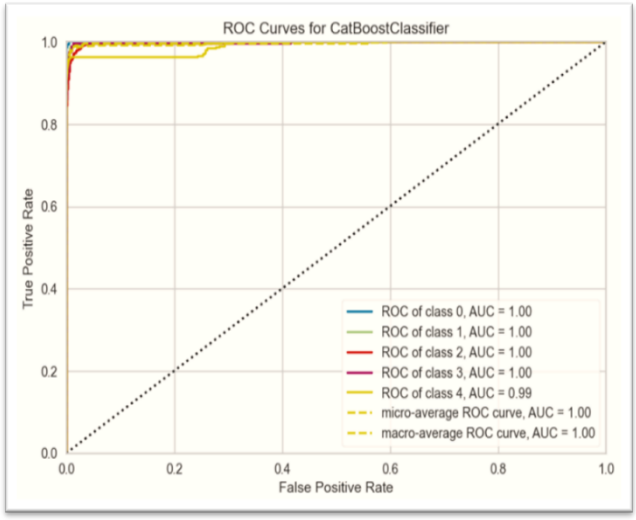


**Figure 16:** Indian Dataset ROC Curves of CatBoost Classifier.

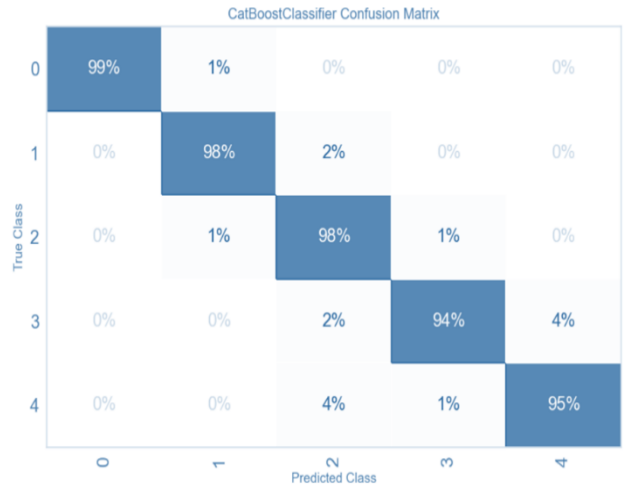


**Figure 17:** Indian Dataset Confusion Matrix of CatBoost Classifier.

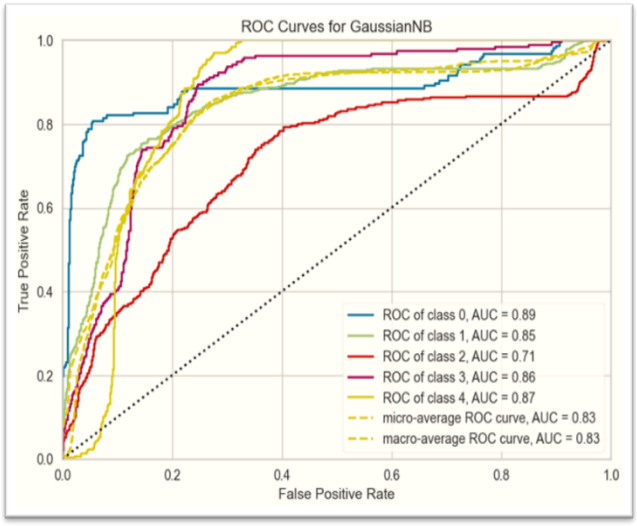


**Figure 18:** Indian Dataset ROC Curves of Naïve Bayes Classifier.

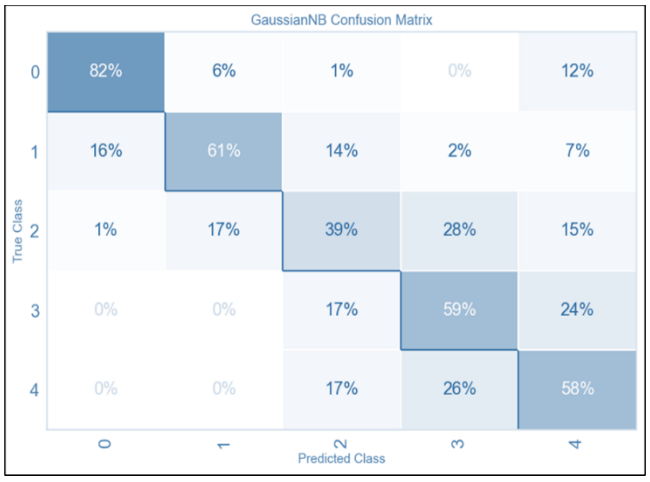


**Figure 19:** Indian Dataset Confusion Matrix of Naïve Bayes Classifier.

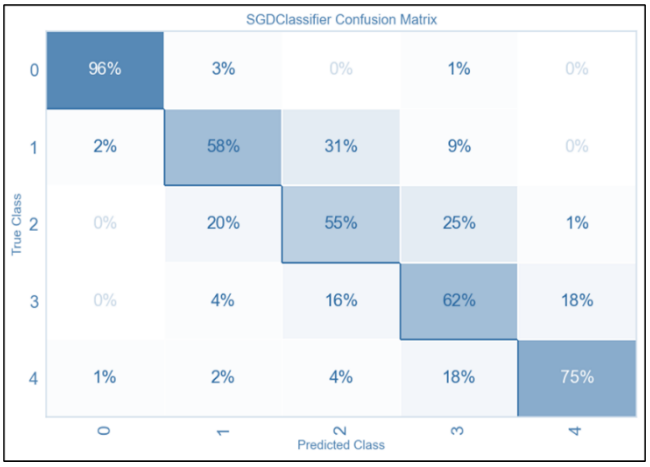


**Figure 20:** Indian Dataset Confusion Matrix of SVM Classifier.

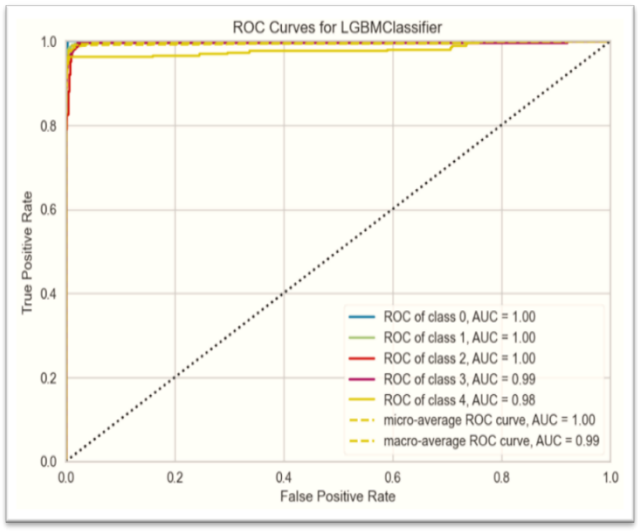


**Figure 21:** Indian Dataset ROC Curves of LGBM Classifier.

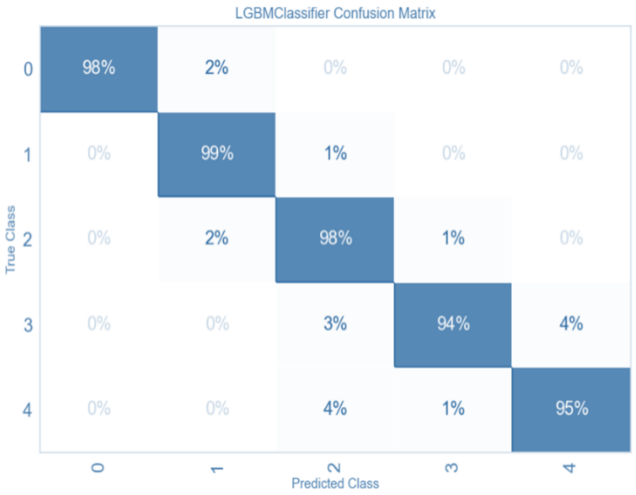


**Figure 22:** Indian Dataset Confusion Matrix of LGBM Classifier.

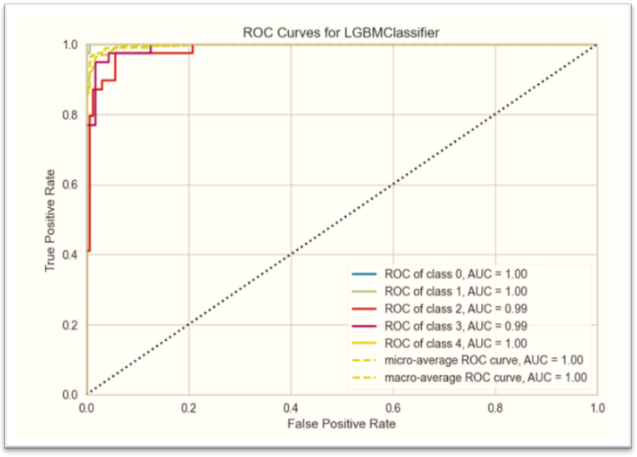


**Figure 23:** Malaysian Dataset ROC Curves of LGBM Classifier.

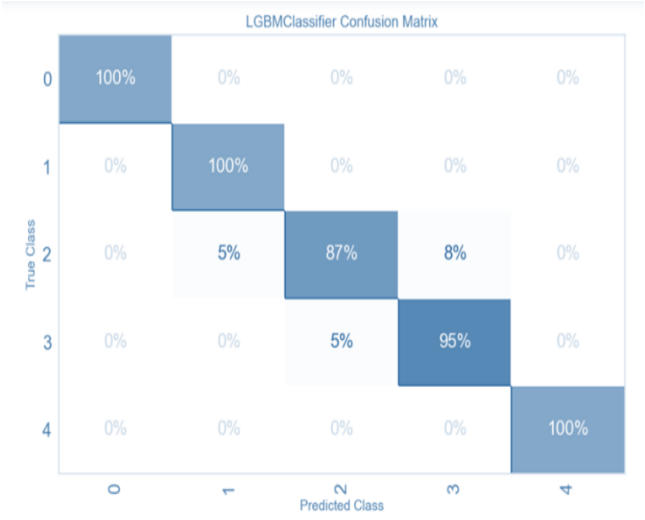


**Figure 24:** Malaysian Dataset Confusion Matrix of LGBM Classifier.

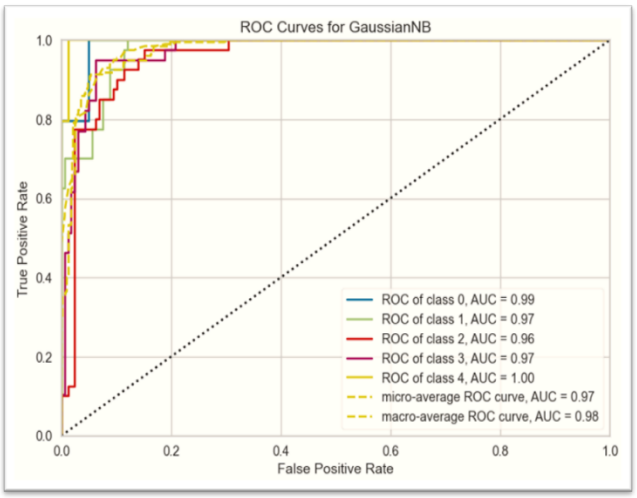


**Figure 25:** Malaysian Dataset ROC Curves of Naïve Bayes Classifier.

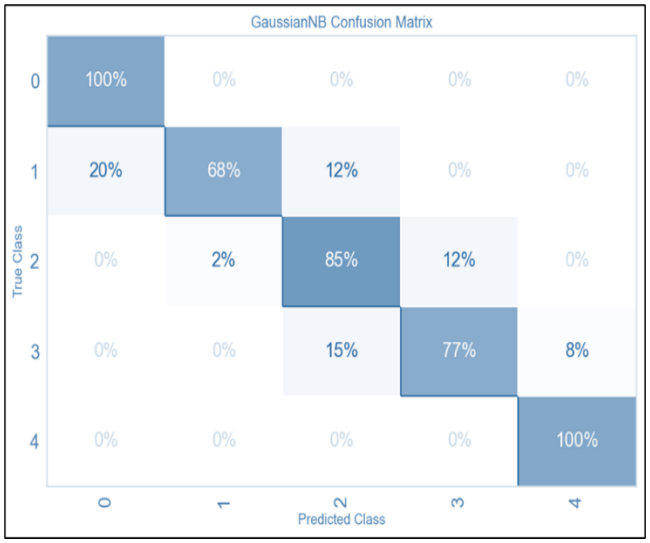


**Figure 26:** Malaysian Dataset Confusion Matrix of Naïve Bayes Classifier.

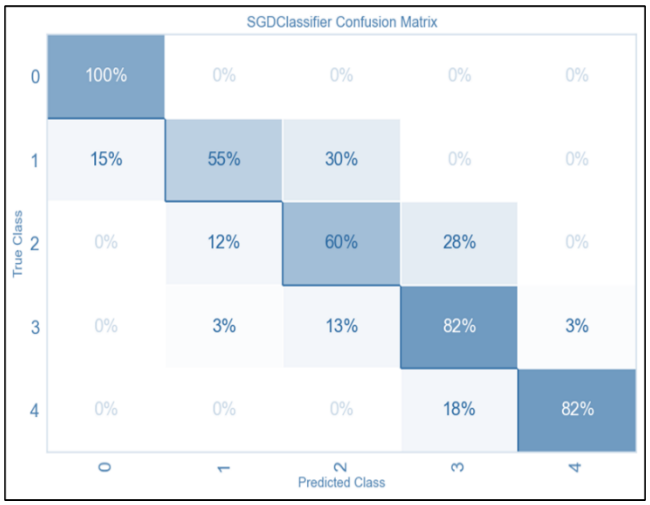


**Figure 27:** Malaysian Dataset Confusion Matrix of SVM Classifier.

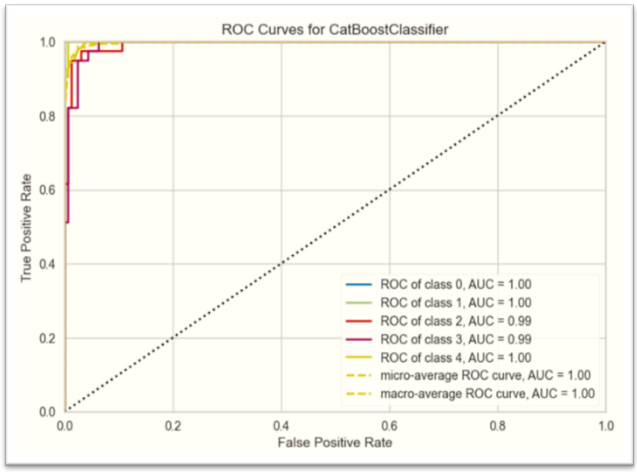


**Figure 28:** Malaysian Dataset ROC Curves of CatBoost Classifier.

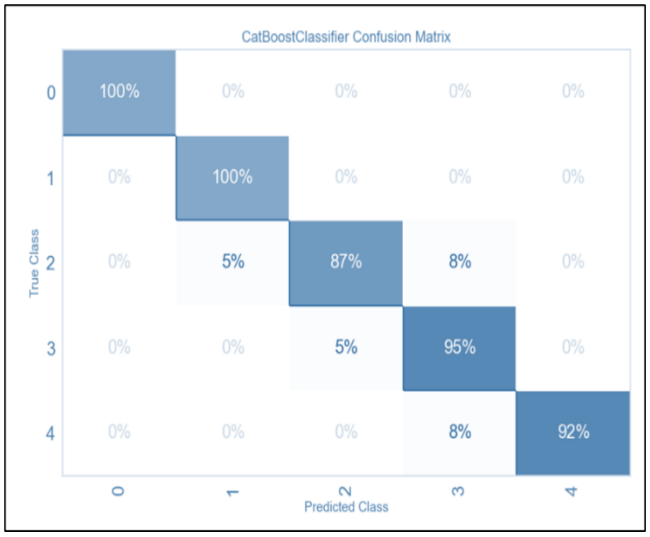


**Figure 29:** Malaysian Dataset Confusion Matrix of CatBoost Classifier.

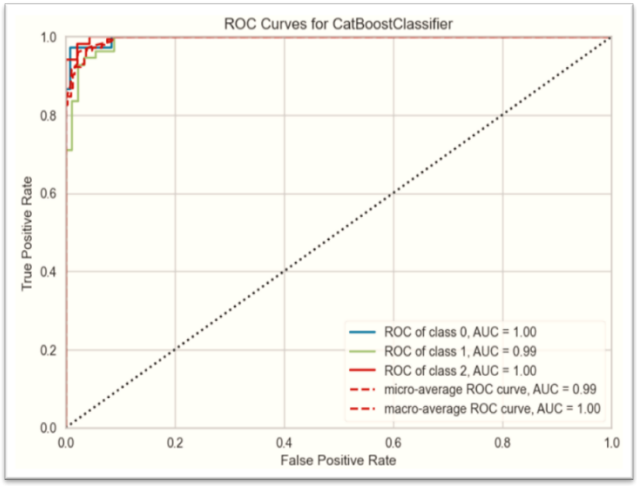


**Figure 30:** Iraq Dataset ROC Curves of CatBoost Classifier.

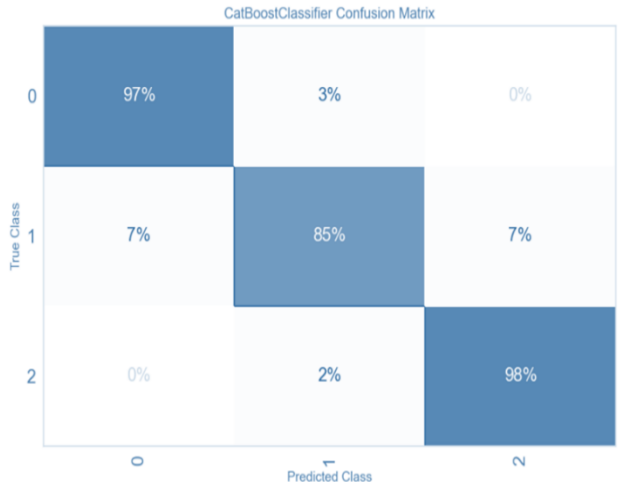


**Figure 31:** Iraq Dataset Confusion Matrix of CatBoost Classifier.

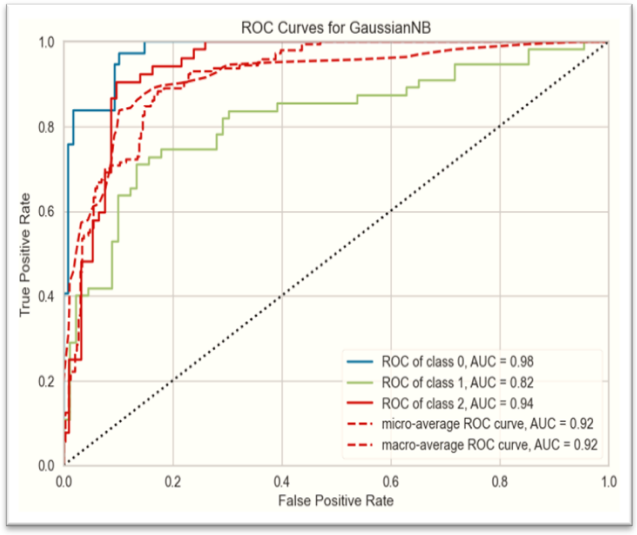


**Figure 32:** Iraq Dataset ROC Curves of Naïve Bayes.

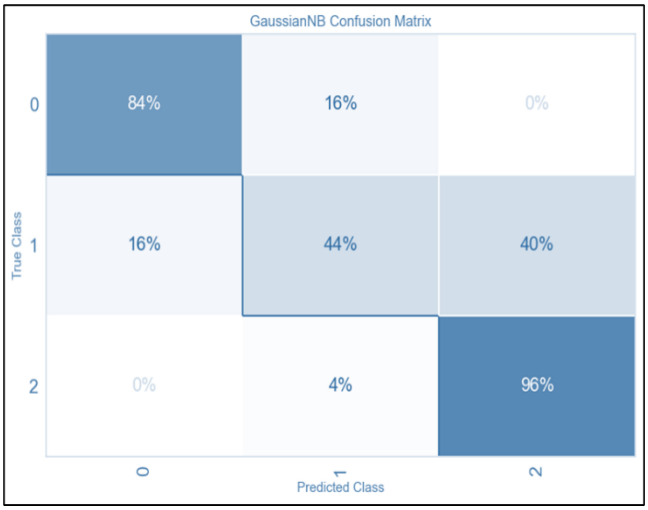


**Figure 33:** Iraq Dataset Confusion Matrix of of Naïve Bayes.

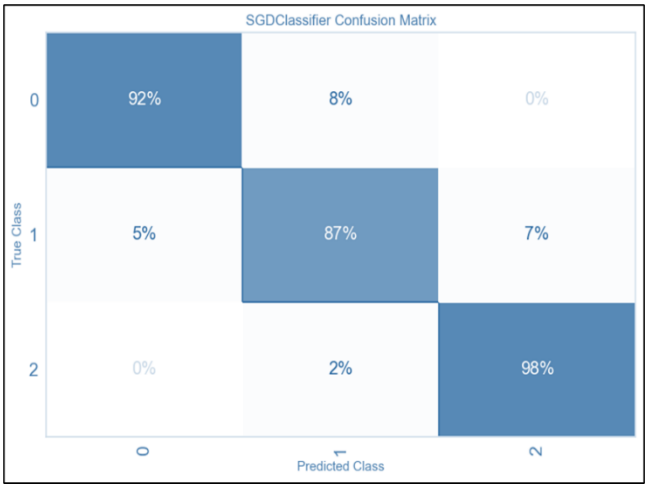


**Figure 34:** Iraq Dataset Confusion Matrix of SVM Classifier.

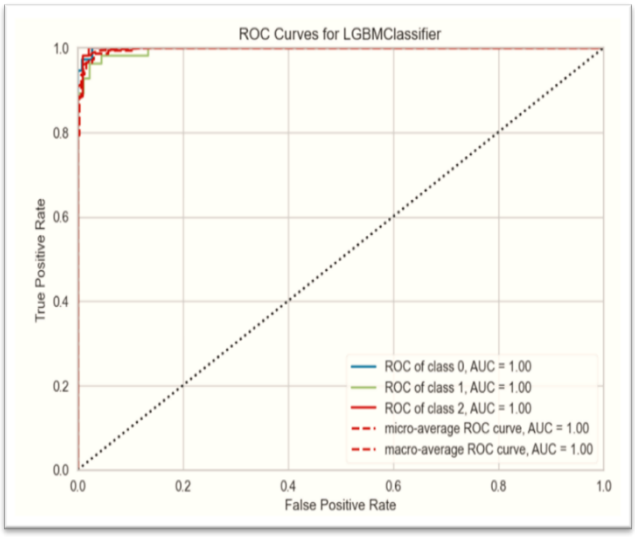


**Figure 35:** Iraq Dataset ROC Curves of LGBM Classifier.

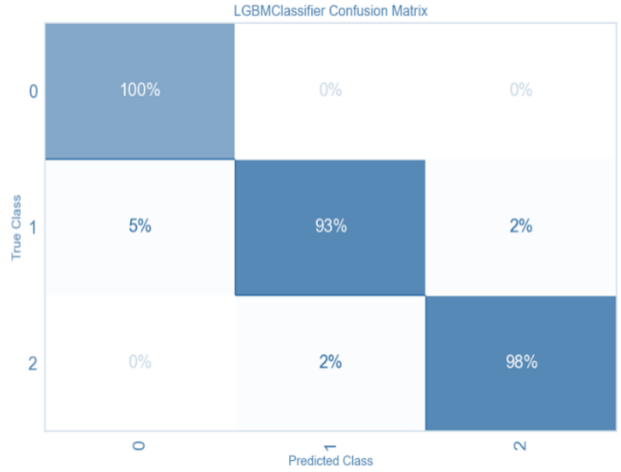


**Figure 36:** Iraq Dataset Confusion Matrix of LGBM Classifier.