



Water quality prediction using machine learning models based on grid search method

Mahmoud Y. Shams¹ · Ahmed M. Elshewey² · El-Sayed M. El-kenawy³ · Abdelhameed Ibrahim⁴ · Fatma M. Talaat^{1,5} · Zahraa Tarek⁶

Received: 17 June 2022 / Revised: 10 August 2023 / Accepted: 31 August 2023 /

Published online: 29 September 2023

© The Author(s) 2023

Abstract

Water quality is very dominant for humans, animals, plants, industries, and the environment. In the last decades, the quality of water has been impacted by contamination and pollution. In this paper, the challenge is to anticipate Water Quality Index (WQI) and Water Quality Classification (WQC), such that WQI is a vital indicator for water validity. In this study, parameters optimization and tuning are utilized to improve the accuracy of several machine learning models, where the machine learning techniques are utilized for the process of predicting WQI and WQC. Grid search is a vital method used for optimizing and tuning the parameters for four classification models and also, for optimizing and tuning the parameters for four regression models. Random forest (RF) model, Extreme Gradient Boosting (Xgboost) model, Gradient Boosting (GB) model, and Adaptive Boosting (AdaBoost) model are used as classification models for predicting WQC. K-nearest neighbor (KNN) regressor model, decision tree (DT) regressor model, support vector regressor (SVR) model, and multi-layer perceptron (MLP) regressor model are used as regression models for predicting WQI. In addition, preprocessing step including, data imputation (mean imputation) and data normalization were performed to fit the data and make it convenient for any further processing. The dataset used in this study includes 7 features and 1991 instances. To examine the efficacy of the classification approaches, five assessment metrics were computed: accuracy, recall, precision, Matthews's Correlation Coefficient (MCC), and F1 score. To assess the effectiveness of the regression models, four assessment metrics were computed: Mean Absolute Error (MAE), Median Absolute Error (MedAE), Mean Square Error (MSE), and coefficient of determination (R^2). In terms of classification, the testing findings showed that the GB model produced the best results, with an accuracy of 99.50% when predicting WQC values. According to the experimental results, the MLP regressor model outperformed other models in regression and achieved an R^2 value of 99.8% while predicting WQI values.

Keywords Water quality · Machine learning models · Grid search · Water quality index · Water quality classification

1 Introduction

Water is among the most precious resources on which all existence is dependent. Water contamination degrades water quality, impacting the health of sea creatures and, by extension, humans that use them. This makes it critical to observe water quality and ensure the survival of nautical life [1]. Comprehension of water quality concerns and issues is also crucial for water pollution mitigation and control. To grasp the condition of the nautical ecosystem, several governments throughout the world have begun to build ecological water management programs. Roughly one billion individuals do not have access to clean water for drinking, and two million individuals perish every year as a consequence of polluted water and poor sanitation and cleanliness. As a result, preserving the freshwater quality is critical [2]. Water quality is critical to the long-term viability of a diversion plan. The water of poor quality may also be costly since resources must be shifted to repair water delivery infrastructure whenever an issue emerges. The demand for enhanced water management and water quality control has been rising for these objectives to assure safe drinking water at reasonable costs. To address these issues, systematic assessments of freshwater, disposal systems, and organizational monitoring issues are necessary [3]. Forecasting water quality entails anticipating fluctuation characteristics in a water system's health at a specific moment. Assessment of water quality is critical for water quality planning and regulation. Water pollution avoidance and regulation methods may be improved by forecasting future updates in water cleanliness at varying degrees of pollution and designing reasonable water pollution prevention and control techniques. The overall consistency of water should be assessed in water diversion plans. To handle everyday drinking difficulties, a considerable quantity of water is carried. Thus, in today's civilization, solutions for anticipating water quality should be researched [4]. The use of artificial intelligence (AI) and machine learning (ML) technologies is currently critical to security threats [5] and focus on mapping the connection between system inputs and outcomes rather than complex operations strategies [6].

Water quality forecasting is an essential method for water planning, regulation, and monitoring; it is a necessary component of water contamination research to investigate water ecological protection. As a consequence, it is crucial to enhance a realistic and practical strategy for predicting water quality. Simultaneously, forecasting futurity water quality is necessary for preventing sudden updates in water quality and offering solutions. As a result, precise forecast of water quality updates may not only assure the health of individual's potable water but can also help guide fishing productivity and safeguard biodiversity [7]. Furthermore, the typical water quality forecast technique cannot account for the effects of biology, physics, hydraulics, alchemy, and meteorology. At the moment, researchers are primarily concerned with enhancing the practicability and trustworthiness of groundwater forecasting techniques and have presented a range of new techniques, such as artificial neural networks (ANN), stochastic mathematics, fuzzy mathematics, 3S technology, and others, for enhancing water quality forecasting techniques and expand the range of applications [8].

The emergence of remote sensing (RS), cloud computing, the Internet of Things (IoT), big data, and artificial intelligence has created new possibilities for improving and implementing water environment surveillance technologies. Intelligent detection methods for water environmental conservation have been developed in counties and cities throughout China, relying on various types of Stations for spontaneous hydrological and water quality surveillance, wireless sensor networks (WSNs), RS surveillance systems,

surveillance ships, and sophisticated underwater robotic machines [9]. Artificial intelligence solutions may significantly reduce water supply and sanitation systems while also assisting in ensuring acquiescence with consuming water and wastewater handling quality standards. As a result, modeling and forecasting water quality to control water contamination has received a lot of attention [10].

A Water Quality Index (WQI) is a metric utilized to quantify water quality for a variety of reasons. WQI may be used to determine if water is acceptable for consumption, industrial usage, aquatic creatures, etc. The larger the WQI, the higher the water quality [11]. The Water Quality Classification (WQC), which categorizes water as either mildly contaminated or clean, was developed using the WQI value scope [12]. The Water Quality Index (WQI) covers many water quality characteristics at a given location, and time. When doing subindex computations, WQI computation requires time and is frequently influenced by mistakes. As a result, providing an efficient WQI forecasting technique is critical [13].

The extremely nonlinear connections for the researched system can be correctly modeled with or without previous information through gaining knowledge from a large amount of historical data that incorporates the dynamic development operation [14].

Clean water is a crucial item on which living organisms rely. As a result, developing a water quality forecasting technique to forecast futurity water quality situation has enormous gregarious and economic significance [7].

Water quality has been greatly impacted by contamination and pollution in recent decades, which has had a negative impact on both aquatic ecosystems and human health. Understanding and analysing water quality is critical to guaranteeing the long-term usage and management of this valuable resource. The Water Quality Index (WQI) is a well recognised indicator that gives a thorough assessment of water quality based on various parameters. It gives a quantitative metric that reduces the complicated nature of water quality into a single number, allowing for easy interpretation and comparison across multiple sites and time periods. WQI considers a variety of physical, chemical, and biological characteristics such as pH, dissolved oxygen, turbidity, nutrient levels, and the presence of pollutants. WQI gives a thorough evaluation of water quality by aggregating these factors, which supports in decision-making processes linked to water resource management. Water quality grading (WQC) is an additional feature that categorises water samples into specified quality classes based on predefined thresholds. This categorization gives a realistic framework for determining the amount of pollution in water, allowing for targeted actions and regulatory measures. Stakeholders can identify locations or causes of concern, prioritise remediation activities, and adopt necessary actions to safeguard water resources by grading water quality. The study was motivated by the urgent need to address water quality degradation and its effects. Water pollution and contamination pose serious dangers to ecosystems, public health, and long-term development. Water quality monitoring and assessment are essential steps in recognising possible concerns, adopting effective management plans, and maintaining the supply of clean and safe water for diverse sectors. Traditional techniques of water quality evaluation, which include laboratory analysis and WQI computation utilising measurable parameters, can be time consuming, costly, and restricted in their capacity to offer real-time information. Predictive modelling provides an alternate method by estimating WQI and WQC based on existing data using machine learning techniques. Water quality may be assessed in a timely way by constructing accurate and effective prediction models, even when direct measurement of all parameters is not possible or practicable. For various reasons, predicting WQI and WQC using machine learning models is critical for assessing water suitability:

Just-in-time water quality monitoring: Predictive models allow for real-time or near-real-time estimate of WQI and WQC, which is more efficient and cost-effective than standard laboratory analysis. This capacity enables continuous water quality monitoring, early identification of degradation, and prompt reaction to possible threats or pollution occurrences. **Partial data handling:** Some metrics in water quality monitoring may have missing or incomplete data. Predictive models may cope with such scenarios well by leveraging the existing data and predicting missing values, guaranteeing WQI is calculated even when the entire data set cannot be accessible directly. **Resource optimization:** With more precise WQI and WQC predictions, resources may be allocated more effectively. Decision makers can prioritize sampling efforts, direct monitoring activities to areas of interest, and optimize treatment strategies based on expected water quality classes. **Early Warning Systems:** Predictive models can serve as the basis for developing early warning systems for water quality issues. Through continuous monitoring and forecasting of the Water Quality Index and WQC, potential risks or deterioration in water quality can be identified in advance, enabling proactive measures to be taken to mitigate impacts and protect water resources.

Machine learning algorithms are used in this work to predict water quality index (WQI) and water quality classification (WQC). Grid search is a vital method used for optimizing and tuning the parameters for four classification models, namely the random forest (RF) model, extreme gradient boosting (XGBoost) model, gradient boosting (GB) model, and adaptive boosting (AdaBoost) for predicting WQC, and four regression models, namely K-nearest neighbor (KNN) regressor model, decision tree (DT) regressor model, support vector regressor (SVR) model, and multi-layer perceptron (MLP) regressor model for predicting WQI. In classification, the experimental results illustrated that the GB algorithm attained the greatest results with accuracy equals to 99.5% while predicting WQC values. In regression, the experimental results illustrated that the MLP regressor technique attained the greatest results with R^2 equals 99.8% while predicting WQI values. This paper's contributions are as follows:

- Data preprocessing is applied, including data imputation (mean imputation), and data normalization was performed to fit the data and make it convenient for any further processing.
- grid search is used for optimizing and tuning the parameters for four classification models to predict WQC, and four regression models to predict WQI.
- To assess the performance of the classification techniques, MCC, accuracy, recall, precision, and F1 score were computed, and four evaluation metrics, MAE, MedAE, square MSE, and coefficient of determination (R^2) were computed to evaluate the achievements of the regression models.
- The findings showed that the GB model performed the best in terms of predicting WQC in classification. Furthermore, the experimental findings demonstrated that the MLP regressor model performed the best in terms of predicting WQI in regression.

The remainder of the paper is organized as follows: Section 2 provides some studies related to water quality prediction. Recommended materials and methods in this paper are presented in Section 3. The proposed methodology of our work is illustrated in Section 4. Section 5 shows results and discussion. Finally, the conclusion is summarized in Section 6.

2 Related work

Artificial Neural Networks (ANN), Support Vector Regressions (SVR), Grey Systems (GS), Regression Analyses (RA), and other approaches are commonly used to estimate water quality [3]. Liu et al. [9] predicted the Yangtze River Basin's drinking water quality utilising a long short-term memory (LSTM) network. Dissolved oxygen (DO), pH, chemical oxygen demand (COD), and NH₃-N were used to construct the LSTM algorithm. The LSTM technique has proved potential for surveillance water quality.

Sakshi Khullar and Nanhey Singh [15] presented a Bi-LSTM model based on deep learning (DLBL-WQA) to anticipate the water quality variables of the Yamuna River in India. A comparison showed that the suggested approach surpassed all other approaches in terms of error rates and prediction accuracy. Sani Abba et al. [16] examined four machine learning techniques Neuro-Fuzzy Inference (ANFIS), Backpropagation (BPNN), Multi-layer Perceptron (MLP), and Support Vector Regressor (SVR) for anticipating the water quality index (WQI). The acquired findings demonstrated the viability of the built smart techniques for forecasting the WQI at the three stations using the neural network ensemble's better modeling outcomes (NNE). The predictive comparison indicated that NNE was successful and hence may be used as a trustworthy prediction strategy.

Elbeltagi et al. [17] used four standalone techniques: M5P tree model (M5P), additive regression (AR), support vector machine (SVM), and random subspace (RSS) to forecast WQI depending on variable elimination strategy. AR surpassed each other data-driven approaches. The AR is offered as an optimal approach with good outcomes due to improved forecasting reliability with the fewest source variables and could thus be used to anticipate WQI in the Akot basin dependably and exactly. Seyed Asadollah et al. [18] presented Extra Tree Regression (ETR), an ensemble machine learning technique, for forecasting monthly WQI rates along the Lam Tsuen River in Hong Kong. The results of the comparison between ETR and conventional standalone approaches (SVR, DTR), revealed that the ETR approach delivers superior reliable WQI forecasts in both the training and testing stages. Generally, the ETR approach outperformed earlier techniques for WQI forecasting in terms of predictive accuracy and the number of input variables. Moreover Nosair 2022 et al. [19] presents a predictive regression model based on an original strategy employing SWI indicators and artificial intelligence (AI) approaches to monitor groundwater salinization due to saltwater intrusion (SWI) in the aquifer of the eastern Nile Delta, Egypt. Farid Garabaghi et al. [20] presented four machine learning techniques with ensemble learning approaches, namely Random Forest, LogitBoost, XGBoost, and AdaBoost for categorization of the water quality. As a consequence, XGBoost outperformed the other classification methods, with an accuracy of 96.9696 percent when important characteristics were included in the classification stage. The XGBoost model is recommended as the greatest classification method with high accuracy of 95.606 percent with tenfold cross validation When the classification stage involved seven variables selected by the Backward Feature Elimination Feature selector. Mehedi Hassan et al. [21] applied machine learning algorithms such as NN, RF, SVM, BTM, and MLR to classify a water quality dataset in diverse locations throughout India. Biological oxygen demand (BOD), dissolved oxygen (DO), total coliform (TC), pH, Nitrate, and electric conductivity (EC) are all factors that influence water quality. These characteristics are dealt with in 5 stages: min–max normalization for data pre-processing and missing data maintaining using RF, feature correlation, applied machine learning categorization, and classification significance. This study's maximum accuracy, accuracy upper, kappa, and accuracy lower results are 99.83, 99.99, 99.17, and

99.07, respectively. The results revealed that conductivity, Nitrate, DO, PH, BOD, and TC are the main attributes that help to organize the classification of water quality, with parameter significance results of 81.494, 74.78, 105.770, 36.805, 130.173, and 105.166, respectively. Table 1 lists some of the machine learning models for water quality prediction.

According to the previous works, the prediction and classification accuracy is improved using machine learning techniques, so we discuss the effect of some of the machine learning techniques in the next section to predict water quality in a high percentage for prediction and classification.

3 Materials and methods

Following the primary data preprocessing, a particular ML approach is chosen to be trained and verified using the training and validation sets. Before being tested, the corresponding hyper variables will be fine-tuned until the predetermined training target is satisfied. The test dataset will eventually be applied to evaluate the trained approach and assess its enhancement. For clarity, the ML modeling flow chart is given in Fig. 1. The general block diagram of ML models begins with data splitting and preprocessing, followed by model selection. The selected model then undergoes training, testing, and validation. Cross-validation is used to evaluate whether the training model has met its goals. If so, the model can proceed to testing and performance assessment. If not, the model parameters need further fine-tuning during training. To increase the effectiveness of water quality prediction in this work, eight frequently used ML approaches are refined, implemented, and used, as shown below.

3.1 Classification model for predicting WQC

This section introduced four classification algorithms: RF, XGBoost, GB, and AdaBoost.

3.1.1 Random Forest (RF)

RF method is an ensemble technique used for categorization. It is a supervised machine learning method composed of numerous decision trees. Because it is an ensemble technique, it uses the best outcome given by the many decision trees, mitigating and limiting generalization mistakes as the volume of the tree architecture in the forest grows [26]. The classification and regression tree (CART) algorithm is used by the decision tree to categorize the tuples depending on the target parameter. This approach is applied in conjunction with bagging for resampling goals, updating the training data as a new tree forms [27].

Based on the parameters and equations listed below, a tree structure is built to categorize the features [1]. The Gini Index may be used to create the decision tree for any tuple S and is determined using the formula:

$$Gini(y, s) = 1 - \left(\sum_{c \in \text{dom}(y)} \left(\frac{|\sigma_y = c_j \cdot S|}{|S|} \right)^2 \right) \quad (1)$$

The entropy and information gain are also important when creating a decision tree and determining its outcome. It may be computed using the following formulas:

Table 1 ML techniques for water quality prediction

| Author | Technique | Best Model | Prediction Index | Results |
|------------------------------------|---|-------------------------------------|---|--|
| Radhakrishnan and Pillai [22] | Support Vector Machine, Decision Tree, Naïve Bayes | Decision Tree Algorithm | weighted arithmetic water quality index (WAWQI) | Accuracy = 98.50% |
| Damish Jain et al. [1] | Random Forest Algorithm, SVM, K-Nearest Neighbors (KNN) | Random Forest Algorithm | Water Quality Index (WQI) | Accuracy = 92.127% |
| Hmoud Al-Adhailah and Alsaade [10] | Neuro-Fuzzy Inference (ANFIS), KNN, Feed-forward neural network (FFNN) | ANFIS for (WQI) and FFNN for (WQC) | Water Quality Classification (WQC), Water Quality Index (WQI) | Accuracy(ANFIS) = 96.17% Accuracy(FFNN) = 100% |
| Malek et al. [12] | DT, Naive Bayes, Gradient Boosting, KNN, ANN, RF, SVM | Gradient Boosting | Water Quality Classification (WQC) | Accuracy = 94.90% |
| Khan et al. [23] | Principal Component Regression (PCR), Gradient Boosting Classifier (GBoost) | Gradient Boosting Classifier | Water Quality Index (WQI), Water Quality Status (WQS) | Accuracy (PCR) = 95% Accuracy (GBoost) = 100% |
| Theyazn Aldhyani et al. [24] | Neural Autoregressive Network (NARNET), SVM, KNN, Naïve Bayes, Long Short-Term Memory | NARNET for (WQI) and SVM for (WQC) | WQC (Water Quality Classification), WQI (Water Quality Index) | Accuracy (SVM) = 97.01% R ² (NARNET) = 96.17 |
| Dao Khoi et al. [25] | (Adaptive boosting, GBoost, HGBoost, LGBBoost, XGBoost), (DT, ET, RF), (MLP, RBF, DFFNN, CNN) | Extreme gradient boosting (XGBoost) | WQI | R2 = 0.989 and RMSE = 0.107 |

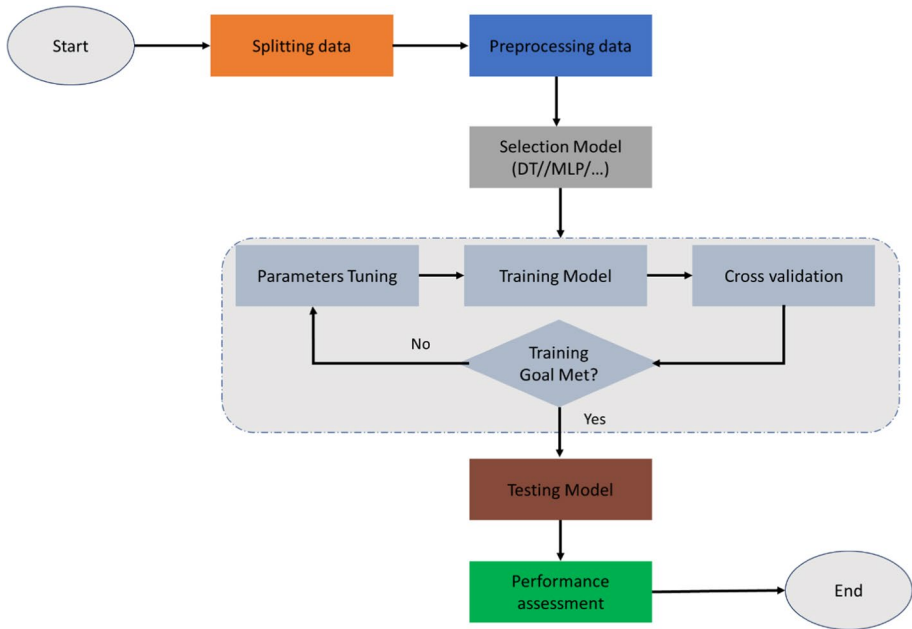


Fig. 1 The flow chart of general machine learning modeling

$$Entropy(S) = \sum -p(i)\log_2 p(i) \tag{2}$$

where p is the fraction of S that belongs to class ‘i’, for each given set S.

$$Gain(S, A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v) \tag{3}$$

where Sv denotes the subset of S for which parameter A has value v.

RF presents numerous benefits. It avoids the issue of multivariate collinearity, which is a disadvantage of ordinary regression analysis. It excels in regression and classification and has a solid grasp of multi-dimensional data [28].

3.1.2 Extreme Gradient Boosting (XGBoost)

The XGBoost is a decision tree enhancement approach that is distinct from the classic gradient boosting decision tree methodology [29]. Based on the optimization issue, the standard GBDT solely employs first-order derivative information. The loss function is then subjected to the second Taylor extension, which employs the first and second-order derivatives. The loss function includes a regularization term to manage the technique’s intricacy and prevent overfitting. The XGBoost technique is derived as follows [28]:

$$\hat{y}_i = \phi(X_i) = \sum_{k=1}^K f_k(X_i), f_k \in F \tag{4}$$

where $F = \{f(x) = w_{q(x)}\} (q : R^m \rightarrow T, w \in R^T)$ indicates a function space that defines a decision tree and T is the leaf nodes number of a decision tree. The following is the loss function:

$$L(\phi) = \sum_i I(y_i; y_i) + \sum_k \Omega(f_k) \quad (5)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (6)$$

The first component in Eq. (5) presents the number of leaves, while the second component is the size of the outcome. XGBoost calculates Gain for every node in the tree to assess whether the generated branch is relevant.

$$\text{Gain} = \frac{1}{2} (\text{Gain}_L + \text{Gain}_R - \text{Gain}_O) - \gamma \quad (7)$$

where Gain_O denotes the authentic gain before splitting and $-\gamma$ is the number of the new leaves.

3.1.3 Gradient Boosting (GB) model

The GB is a Machine Learning approach that combines many weak classification methods, often decision trees, to produce a reliable classifier for classification and regression tasks. It builds the system in stages, much like the other boosting strategies, and generalizes it by maximizing an appropriate cost function. In the GB method, improperly identified instances for one step are given more weight in the following step. The benefits of GB include great prediction accuracy and a quick process [30]. This approach is quite identical to Adaptive Boosting (AdaBoost), although AdaBoost has the disadvantage of being greatly impacted by outliers and readily overpowered by noisy data [31].

3.1.4 Adaptive Boosting (Adaboost) model

The AdaBoost method enhances the performance of the classifier by integrating numerous weak learners into a single strong one. It repeatedly adjusts sample weights depending on classification mistakes, raising the weights of misclassified samples while reducing the weights of well-classified samples. As a result, classification methods that focus on miscategorized data rather than minority class examples are used. Because AdaBoost concentrates on prediction performance, the method is biased toward the majority class, which provides more to total prediction performance [32].

3.2 Regression models for predicting WQI

In this section, four regression algorithms, namely, KNN, DT, SVR, and MLP, were presented.

3.2.1 K-Nearest Neighbors (KNN) model

The KNN technique distinguishes samples by locating the nearest neighboring provided points and assigning the majority of n neighbors to a class. If there is a tie, many ways may

be employed to settle it. Nevertheless, KNN is not recommended for big datasets because it does all computation throughout testing and converges during all trained data, calculating the closest neighbor each time [33]. To locate the nearest neighbor in the features vector, the Euclidean distance function (D_i) was used as follows:

$$D_i = \sqrt{(x_1 - x_2) + (y_1 - y_2)^2} \quad (8)$$

where $x_1, x_2, y_1,$ and y_2 are parameters for data input.

3.2.2 Decision Tree (DT)

The DT is a straightforward, basic approach that generates judgments depending on the values of all relevant input variables. DT chooses the root parameter based on entropy before analyzing the weights of the other variables. DT gathered all variable decisions grouped in a top-down tree and prepares the choice based on various values from special attributes. Previous research has revealed that decision tree models work well on unbalanced data. Nevertheless, ensemble techniques based on decision trees, such as Gradient Boosting (GB) and Random Forest (RF), virtually usually surpass single decision trees [12]. The benefits of decision-tree-based models are their insensitivity to missing values, ability to maintain both regular qualities and data, and high efficiency. Decision-tree-based techniques, as compared to other ML algorithms, are better for short-term forecasting and may have a faster computation speed [34].

3.2.3 Support Vector Regression (SVR)

The SVR is a machine learning technique that originated from the SVM and is seen to be a promising method for solving nonlinear issues such as regression, forecasting, categorization, and function estimation. The technique is an effective method for resolving convex quadratic programming issues. Furthermore, SVR has outstanding characteristics such as non-convergence to a local optimum, a strong mathematical formulation, great predictability, and scalability. Nevertheless, the training dataset must be manually annotated, and the SVR technique's three variables must be changed using prior information [35–37]. SVR's generic nonlinear function is as follows:

$$y(x) = W^T \varphi(x) + b \quad (9)$$

where y represents the link between predictand and predictors, W denotes the weight vector, $\varphi(x)$ is the input dataset's nonlinear mapping function, and b presents the scalar threshold. Figure 2 depicts the SVR structure.

3.2.4 Multi-Layer Perceptron (MLP) regressor

The MLP has an input–output layers and numerous hidden layers. The source signal is transferred forward through the input layer to the hidden layer, where the neurons are computationally managed before being provided forward to the output layer. The output of the MLP neural network depends only on the current input and not on preceding or future inputs; as a result, the MLP neural network is also referred to as a multi feed-forward neural network. MLP neural networks are among the numerous neural network designs that are basic in framework, simple to execute, and have strong fault tolerance, resilience,

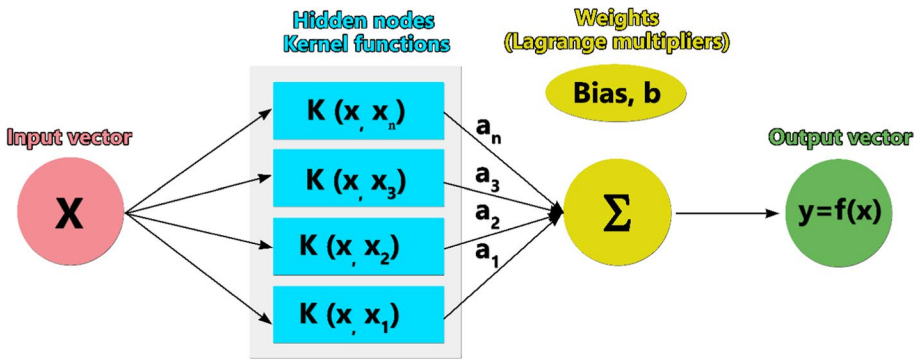


Fig. 2 Structure of the SVR model

scalability, and outstanding nonlinear mapping capabilities [7]. Figure 3 depicts the architecture of the MLP neural network.

4 Proposed methodology

Water contamination is one of the most serious environmental issues confronting humanity, and the damage it causes is mostly due to a lack of forecasting, early caution, and emergency management capabilities. As a result, the implementation of an appropriate surveillance and early alert system to enable intelligent decision making and water quality management is a critical scientific and technical issue that must be addressed promptly [38]. Several machine learning approaches have advanced rapidly in recent years, Fig. 4 shows the proposed methodology to predict the quality of water.

The proposed methodology aims to develop a machine learning model for water quality assessment based on a dataset containing seven features: dissolved oxygen, pH, conductivity, biological oxygen demand, nitrate, fecal coliform, and total coliform. The dataset has already undergone preprocessing, which includes mean imputation and data normalization.

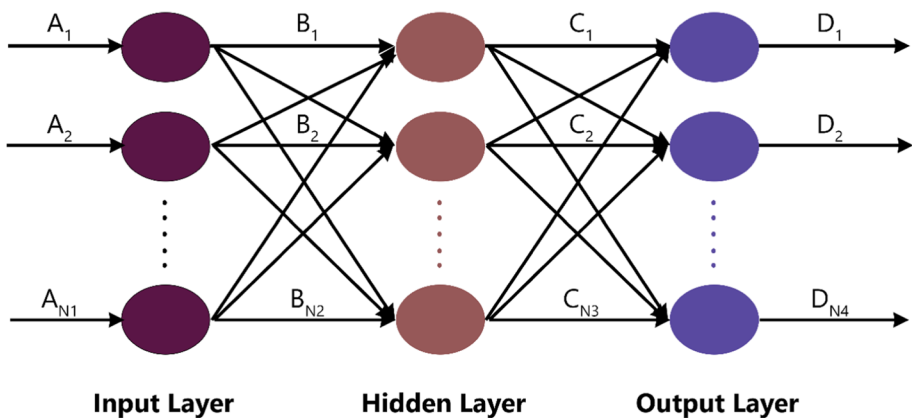


Fig. 3 MLP neural network topology

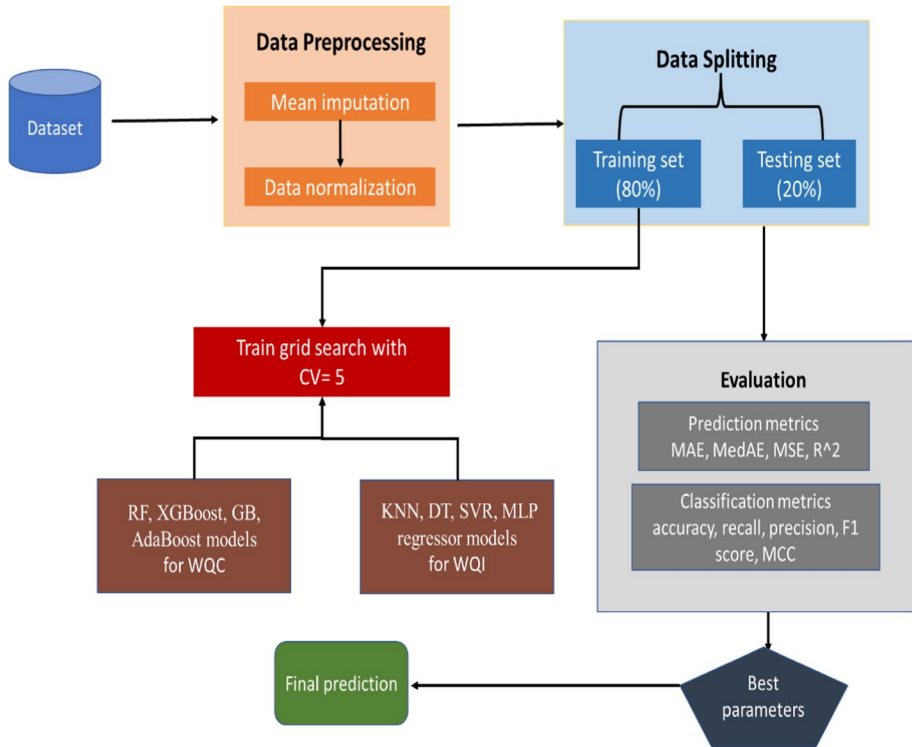


Fig. 4 The proposed methodology

The data has been split into a training set (80%) and a testing set (20%). During the training phase, a grid search with cross-validation ($CV = 5$) is used to tune hyperparameters for four different models for water quality classification (RF, XGBoost, GB, and Adaboost) and four different models for water quality index (KNN, DT, SVM, and MLP).

The features of the data, the problem being handled, and the application's performance requirements all influence the choice of certain classification and regression models. The specific models used in the Water Quality Assessment method were most likely chosen based on their ability to handle the features of the water quality dataset and their performance on similar situations. The presented ensemble models combine numerous weak learners to create a stronger model. These models are frequently employed in classification problems with a high number of characteristics and complicated interactions between the variables and the target variable in the dataset. Ensemble approaches can capture these complicated interactions and increase model accuracy. RF is well-known for its capacity to handle high-dimensional data while avoiding overfitting, whereas Xgboost, GB, and AdaBoost are well-known for their rapid training and prediction times as well as excellent accuracy.

Popular regression models include KNN, DT, SVM, and MLP, which can handle diverse types of data and correlations between features and the target variable. The KNN model is a non-parametric model that can handle both linear and non-linear correlations between features and the target variable. DT is a tree-based paradigm that can manage non-linear

connections and has a straightforward interpretation. The SVR is a kernel-based model that works well on small datasets and can manage non-linear connections. A MLP is a neural network-based model that can handle complex interactions between features and the target variable.

During the testing phase, the models' performance is evaluated using various metrics such as Mean Absolute Error (MAE), Median Absolute Error (MedAE), Mean Squared Error (MSE), R-squared (R²) for prediction, and accuracy, recall, precision, F1 score, and Matthews Correlation Coefficient (MCC) for classification.

Grid search is a hyperparameter tuning approach often used in machine learning to discover the optimal hyperparameter combination for a given model. Hyperparameters are parameters that must be specified before to training the model and cannot be learnt from data. The learning rate, the regularization parameter, the number of layers in a neural network, and the number of trees in a random forest are all examples of hyperparameters.

Grid search seeks to extensively search through all potential hyperparameter combinations within a particular range or set of values. This is performed by first creating a grid of all possible hyperparameter combinations, and then training and testing the model on a validation or cross-validation set for each combination. The optimal set of hyperparameters is the set of hyperparameters that gives the best performance on the validation or cross-validation set.

The grid search algorithm is explained as follows:

- Define the hyperparameters as well as their potential values or ranges.
- Make a grid with all conceivable hyperparameter combinations.
- For each hyperparameter combination in the grid:
 - a Train the model on the training set using the current hyperparameters.
 - b Using a performance metric, evaluate the model on the validation or cross-validation set (CV = 5).
 - c Keep track of the performance statistic.
- Choose the hyperparameter combination that produced the best performance measure.

Grid search may be computationally costly, particularly when there are a large number of hyperparameters and their possible values or ranges. Using randomized search instead of grid search can help to lower computing costs. A random subset of hyperparameters is sampled in randomized search.

4.1 Dataset

The dataset used for this study is available at <https://www.kaggle.com/datasets/anbarivan/indian-water-quality-data>. The dataset was collected from lakes and rivers in India from several locations in the period between 2005 to 2014. The government of India collected this data to be sure that the water is valid for drinking. The dataset consists of 1991 instances and 7 features. The dataset features are dissolved oxygen, PH, conductivity, biological oxygen, nitrate, fecal coliform, and total coliform. The features of the dataset are *Dissolved Oxygen* by which it indicates the level of oxygen dissolved in the water, which is essential for supporting aquatic life. *pH*: It represents the acidity or alkalinity of the water, indicating its level of acidity or basicity. The *conductivity*

Table 2 Statistical calculation of the features

| | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
|-------------------|-------|-----------|--------------|------|-------|--------|---------|-------|
| Dissolved_oxygen | 1991 | 6.392637 | 1.322515e+00 | 0.0 | 5.95 | 6.70 | 7.2 | 11.4 |
| PH | 1991 | 112.0906 | 1.875150e+03 | 0.0 | 6.9 | 7.30 | 7.7 | 67115 |
| Conductivity | 1991 | 1786.466 | 5.517290e+03 | 0.4 | 79 | 187.63 | 620.5 | 65700 |
| Biological_oxygen | 1991 | 6.940049 | 2.908065e+01 | 0.1 | 1.20 | 1.90 | 3.9 | 534.5 |
| Nitrate | 1991 | 1.623079 | 3.852301e+00 | 0.0 | 0.28 | 0.62 | 1.62307 | 108.7 |
| Fecal_coliform | 1991 | 362,529.3 | 8.038807e+06 | 0.0 | 41 | 313 | 4950.5 | 27252 |
| Total_coliform | 1991 | 533,687.1 | 1.375409e+07 | 0.0 | 118 | 542 | 2929 | 51109 |
| WQI | 1991 | 75.64109 | 1.359473e+01 | 19.3 | 67.38 | 78.74 | 83.7 | 99.8 |

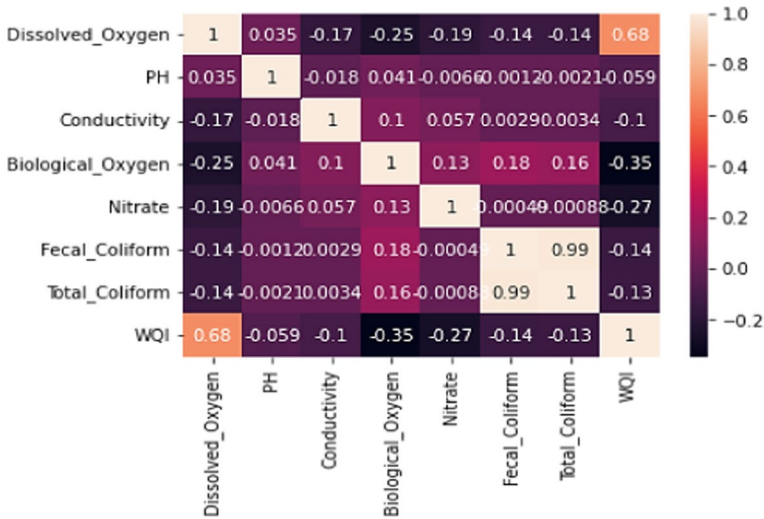


Fig. 5 Heat map visualization of the feature correlations

of water, which evaluates its capacity to conduct electrical current and offers information on the existence of dissolved solids. The *Biological Oxygen Demand (BOD)* is a measurement of the quantity of dissolved oxygen absorbed by microorganisms in water, which indicates the extent of organic contamination. The *Nitrate* that examines the concentration of nitrate ions in water, which can be a sign of fertilizer or sewage pollution. The *Fecal Coliform* is an indication of faecal pollution since it reflects the presence of coliform bacteria in the water. *Total Coliform*, which represents the total amount of coliform bacteria from both faecal and non-faecal sources. Certain preprocessing processes were conducted to assure the dataset’s quality and usability in the study. These processes involve dealing with missing values and outliers, both of which are significant problems in real-world datasets. The specifics of the data pretreatment stages are not stated in the context supplied. In addition, as shown in Table 2, the study included statistical computations on the dataset attributes. These computations may include metrics such as mean, standard deviation, minimum, maximum, and quartiles, which provide information about the data’s distribution and properties. Furthermore, the correlation

matrix of the dataset features was analyzed, as depicted in Fig. 5. The correlation matrix explores the relationships between the different features, helping identify any significant associations or dependencies among the variables.

4.2 Water Quality Index (WQI) computation

Water quality index (WQI) is a dominant indicator that impact the water quality [39]. WQI is computed via utilizing various parameters. WQI is computed using Eq. (10):

$$WQI = \frac{\sum_{i=1}^N q_i \times w_i}{\sum_{i=1}^N w_i} \tag{10}$$

where N represents the number of the parameters, q_i represents the quality rating scale for the parameter i , and w_i represents the unit weight for the parameter i . q_i is computed using Eq. (11):

$$q_i = 100 \times \left(\frac{v_i - v_{id}}{s_i - v_{id}} \right) \tag{11}$$

where v_i represents the estimated value for the parameter i , v_{id} represents an ideal value for the parameter i while the water is pure, and s_i represents a standard value for the parameter i . The unit weight w_i is computed using Eq. (12):

$$w_i = \frac{k}{s_i} \tag{12}$$

where k represents the constant of proportionality and computed using Eq. (13):

$$k = \frac{1}{\sum_{i=1}^N s_i} \tag{13}$$

Figure 6 demonstrates the distribution of calculated feature (WQI). The statistical calculation for the feature (WQI) is demonstrated in Table 1.

Table 3 demonstrates the unit weight of the features and Table 4 represents the WQC.

Fig. 6 Distribution of calculated WQI

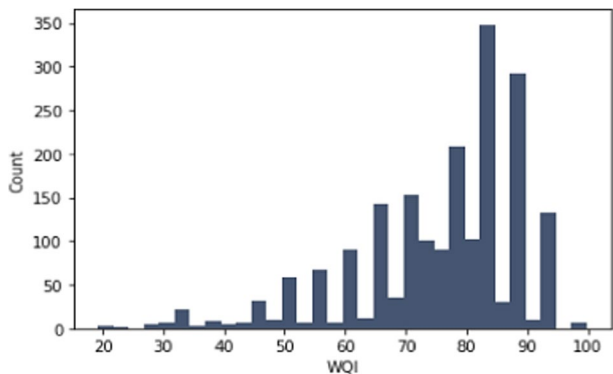


Table 3 Features unit weight

| Features Name | Unit Weight |
|-------------------|-------------|
| Dissolved_oxygen | 0.2213 |
| PH | 0.2604 |
| Conductivity | 0.0022 |
| Biological_oxygen | 0.4426 |
| Nitrate | 0.0492 |
| Fecal_coliform | 0.0221 |
| Total_coliform | 0.0022 |

Table 4 Water quality classification (WQC)

| WQI Rate | Classification |
|---------------|----------------|
| 0–50 | Good |
| 51–100 | Poor |
| More than 100 | Unsuitable |

5 Results and discussion

The experiments are carried out using the jupyter notebook version (6.4.6). Jupyter notebook makes it easier to run and write Python scripts. It is widely used as an open-source model implementation and execution tool for AI and ML. The proposed models' performance is compared to that of numerous existing models. The classification models' performance was assessed using assessment criteria such as accuracy, recall, precision, F1 score, and Matthew's correlation coefficient (MCC). Equation (14) is used to calculate precision:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (14)$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Recall is calculated using Eq. (15):

$$\text{Recall} = \frac{TP}{TP + FN} \quad (15)$$

Precision is calculated using Eq. (16):

$$\text{Precision} = \frac{TP}{TP + FP} \quad (16)$$

F1 score is computed using Eq. (17):

$$\text{F1Score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (17)$$

MCC is calculated using Eq. (18):

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (18)$$

Mean absolute error (MAE), median absolute error (MedAE), mean square error (MSE), and coefficient of determination (R^2) were used to assess the effectiveness of the regression models. Equation (19) is used to compute MAE:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_{real_i} - y_{pred_i}| \tag{19}$$

MedAE is calculated using Eq. (20):

$$MedAE = median\left(|y_{real_1} - y_{pred_1}|, \dots, |y_{real_N} - y_{pred_N}|\right) \tag{20}$$

MSE is calculated using Eq. (21):

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_{real_i} - y_{pred_i})^2 \tag{21}$$

R^2 is calculated using Eq. (22):

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_{real_i} - y_{pred_i})^2}{\sum_{i=1}^N (y_{real_i} - \bar{y})^2} \tag{22}$$

5.1 Water Quality Classification (WQC) prediction

The best parameters for classification models using the grid search approach are shown in Table 5. The table details the tuning parameters investigated for each model, as well as the precise parameter values that resulted in the optimum performance based on the tuning procedure. These best parameters are crucial in optimizing the performance of each machine learning model for their respective tasks. For random forest model, the tuning parameters are:

- N_Estimators that represent the number of decision trees in the forest. The tested values are [50, 100, 150, 200, 250]. The best parameter is 100.
- Criterion that is the function to measure the quality of a split. Tested values are 'gini' and 'entropy'. The best parameter is entropy.

Table 5 The settings of the best parameters for the classification approaches using grid search algorithm

| Approaches | Parameters Tuning | The best parameters |
|------------|--|---|
| RF | Criterion=['gini', 'entropy'] N_Estimators=[50,100,150,200,250], | Criterion=entropy N_Estimators=100, |
| XGBoost | N_Estimators=[50,100,150,200,250], Max_depth=[1,2,3,4,5,6,7,8,9,10], Objective=['binary', 'logistic'] | N_estimators=200, Max_depth=2, Objective=logistic |
| GB | N_estimators=[50,100,150,200,250], Max_depth=[1,2,3,4,5,6,7,8,9,10], Max_features=['auto', 'sqrt', 'log2'] | N_estimators=250, Max_depth=1, Max_features=auto |
| AdaBoost | N_estimators=[50,100,150,200,250], Learning_Rate=[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1] | N_estimators=250, Learning_Rate=0.5 |

For XGBoost model, the tuning parameters are:

- N_Estimators that represent the number of boosting rounds. Tested values are [50, 100, 150, 200, 250]. The best parameter is 200.
- Max_Depth represents the maximum depth of each decision tree. Tested values are [1,2,3,4,5,6,7,8,9,10]. The best parameter is 2.
- Objective is the learning task and corresponding objective. Tested values are 'binary' and 'logistic'. The best parameter is logistic.

For gradient boosting model, the tuning parameters are:

- N_Estimators that is the number of boosting rounds. Tested values are [50, 100, 150, 200, 250]. The best parameter is 250.
- Max_Depth is the maximum depth of each decision tree. Tested values are [1,2,3,4,5,6,7,8,9,10]. The best parameter is 1.
- Max_Features: The number of features to consider when looking for the best split. Tested values are 'auto', 'sqrt', and 'log2'. The best parameter is auto.

For AdaBoost model, the tuning parameters are:

- N_Estimators that represent the maximum number of estimators at which boosting is terminated. Tested values are [50, 100, 150, 200, 250]. The best parameter is 250.
- Learning_Rate that is the rate at which the algorithm adjusts its weights. Tested values are [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1]. The best parameter is 0.5.

Table 6 shows the classification model performance using the grid search strategy.

As shown in Table 6, the performance of the classification models using grid search method, namely, RF model, XGBoost model, AdaBoost model, and the proposed GB model are demonstrated. The results of the proposed GB model demonstrate its superiority over the alternative classification models (highlighted in bold). It achieves an accuracy of 99.5%, F1 score of 99.4%, recall of 99.5%, precision of 99.5%, and Matthews Correlation Coefficient (MCC) of 94.3%. The remarkable performance of the GB model can be attributed to its ability to combine weak learners, specifically decision trees, in an ensemble manner.

Table 7 shows a comparison of the suggested GB classification model utilizing the grid search approach with many research that used the same dataset. The proposed GB model underwent parameter tuning using the grid search method, resulting in exceptional performance. The proposed GB model achieved an impressive accuracy of 99.50% (highlighted in bold). These accuracy values showcase the models' predictive

Table 6 The performance of the classification approaches using the grid search algorithm

| Models | Accuracy | F1 score | Recall | Precision | MCC |
|----------|---------------|---------------|---------------|---------------|---------------|
| RF | 99.00% | 98.90% | 98.90% | 98.90% | 88.50% |
| XGBoost | 99.30% | 99.20% | 99.20% | 99.20% | 91.50% |
| AdaBoost | 99.10% | 99.00% | 99.00% | 99.00% | 88.90% |
| GB | 99.50% | 99.40% | 99.50% | 99.50% | 94.30% |

Table 7 Comparison between proposed GB classification model with several studies used the same dataset

| Studies | Model | Accuracy |
|-------------------|--|---------------|
| Ref [1] | RF | 95.98% |
| Ref [22] | DT | 98.50% |
| Ref [24] | SVM | 97.01% |
| Proposed GB model | Parameters tuning for GB model using grid search | 99.50% |

capabilities, with the Decision Tree model showing higher accuracy than the RF and SVM models. However, the proposed GB model outperformed all other models, achieving the highest accuracy of 99.50%. It is important to note that the GB model’s performance was further enhanced through parameter tuning using the grid search method, showcasing its ability to optimize its predictive accuracy.

Figures 7, 8, 9 and 10 illustrate the feature importance for RF model, XGBoost model, GB model, and Adaboost model, respectively, using grid search method.

Figure 11 shows a comparison between, RF model, AdaBoost model, XGBoost model, and GB model in term of accuracy.

5.2 Water quality index (WQI) prediction

Table 8 shows the best regression model parameters found using the grid search approach. The table summarizes the tuning parameters investigated for each regression model, as well as the exact parameter values that resulted in the best performance during the tuning process. These best parameters play a crucial role in optimizing the models for accurate regression predictions For KNN regressor, the tuning parameters are:

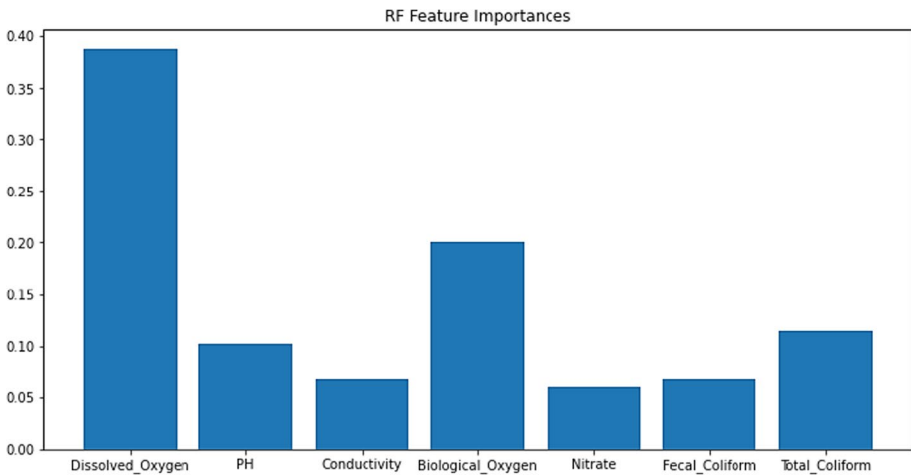


Fig. 7 Feature importance for RF model

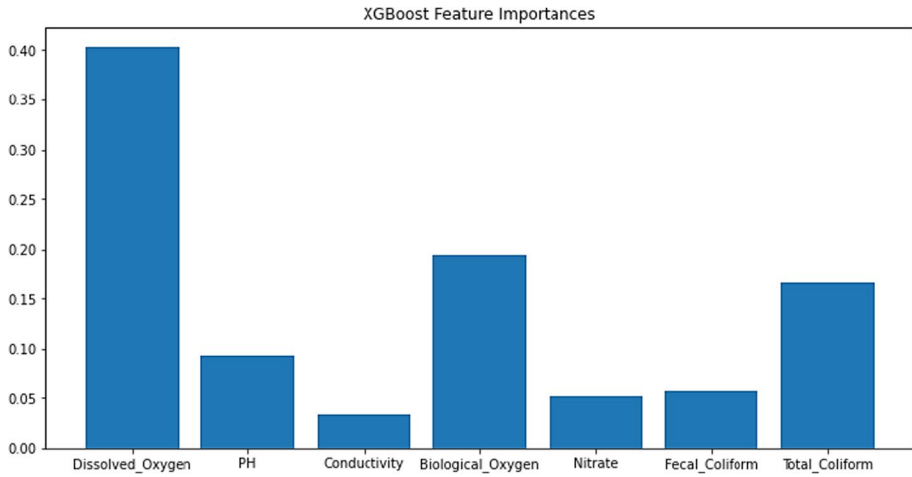


Fig. 8 Feature importance for XGBoost model

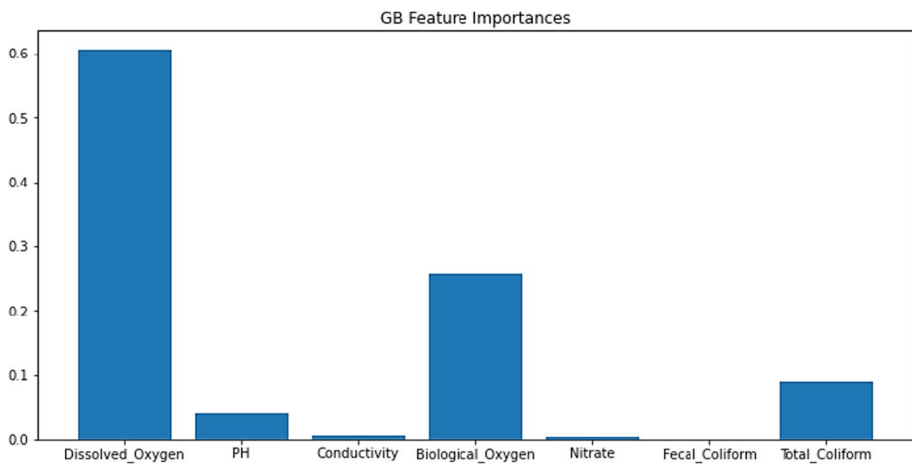


Fig. 9 Feature importance for GB model

- `N_neighbors` represent the number of neighbors to consider for prediction. Tested values are integers from 1 to 50. The best parameter is 1.
- `Weights` is the weight function used in prediction. Tested values are 'uniform' and 'distance'. The best parameter is distance.

For DT regressor, the tuning parameters are:

- `Max_depth` is the maximum depth of the decision tree. Tested values are integers from 1 to 30. The best parameter is 10.
- `Random_state` is the random seed for reproducibility. Tested values are integers from 1 to 50. The best parameter is 33.

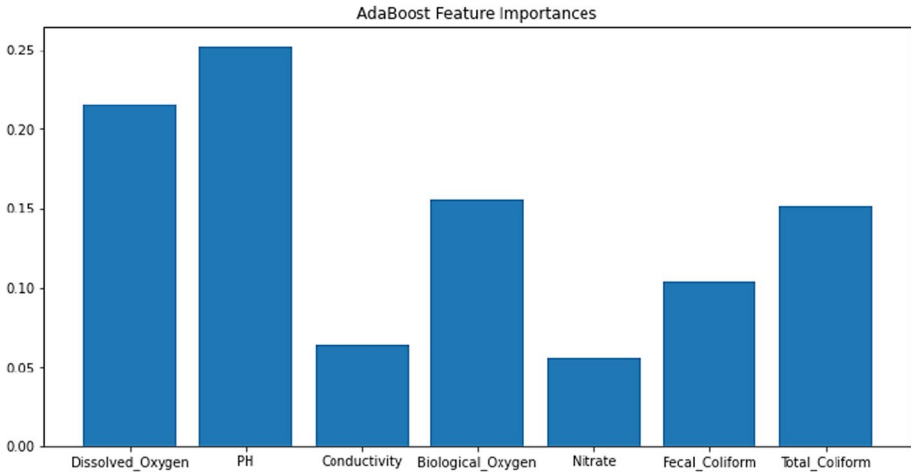


Fig. 10 Feature importances for Adaboost model

Fig. 11 Comparison between, RF model, AdaBoost model, XGBoost model, and GB model in term of accuracy

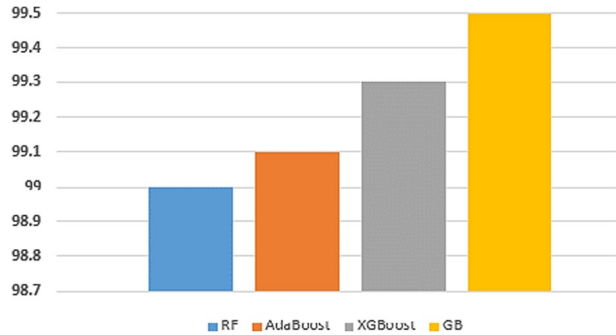


Table 8 Best parameters for the regression models using grid search method

| Models | Tuning parameters | Best parameters |
|---------------|---|---|
| KNN regressor | n_neighbors=[1 to 50], weights=[‘uniform’, ‘distance’] | n_neighbors = 1, weights = distance |
| DT regressor | max_depth=[1 to30], random_state=[1 to 50] | max_depth = 10, random_state = 33 |
| SVR | C=[1,2,3,4,5], epsilon=[0.1, 0.01, 0.001], kernel=[‘sigmoid’, ‘poly’, ‘linear’, ‘rbf’] | C = 2, epsilon = 0.001, kernel = poly |
| MLP regressor | activation=[‘relu’, ‘tanh’, ‘logistic’], solver=[‘sgd’, ‘lbfgs’, ‘adam’], alpha=[0.01, 0.001, 0.0001] | activation = tanh, solver = lbfgs, alpha = 0.0001 |

For SVR regressor model, the tuning parameters are:

- C is the regularization parameter. Tested values are [1, 2, 3, 4, 5]. The best parameter is $C=2$.
- Epsilon is the margin of tolerance for errors. Tested values are [0.1, 0.01, 0.001]. The best parameter is 0.001.
- Kernel is the kernel function used in SVR. Tested values are 'sigmoid', 'poly', 'linear', and 'rbf'. The best parameter is poly.

For MLP regressor model, the tuning parameters are:

- Activation is the activation function in hidden layers. Tested values are 'relu', 'tanh', and 'logistic'. The best parameter is tanh.
- solver is the optimization algorithm. Tested values are 'sgd', 'lbfgs', and 'adam'. The best parameter is lbfgs.
- alpha is the L2 regularization parameter. Tested values are [0.01, 0.001, 0.0001]. The best parameter is 0.0001.

Table 9 describes the performance of the regression models using grid search method.

Table 9 presents the performance of different regression models obtained through the grid search method. These models include the KNN regressor model, DT regressor model, SVR model, and the proposed MLP regressor model. Out of these models, the proposed MLP regressor model achieves the highest performance compared to the other regression models. The performance of the proposed MLP regressor model surpasses the others due to its inherent characteristics and capabilities. One significant advantage of MLP is its ability to learn complex non-linear relationships between the input and output variables. Through a process called backpropagation, the MLP receives feedback on the error in its predictions and adjusts the weights of the connections between neurons to minimize this error. This iterative learning process allows the MLP to continually improve its predictive accuracy. MLP proves to be effective because it can capture and model intricate patterns and dependencies present in the data. By leveraging its hidden layers and the activation functions within them, MLP can approximate complex functions and provide accurate predictions for regression tasks. The results of the proposed MLP regressor model in Table 9 further highlight in bold its superiority over the other regression models. It achieves a Mean Absolute Error (MAE) of 0.003, Mean Squared Error (MSE) of 2.8×10^{-5} , Median Absolute Error (MedAE) of 0.0009, and an R-squared (R^2) value of 99.8%. In contrast, the KNN regressor model demonstrates the lowest performance with an MAE of 0.009, MSE of 0.0002, MedAE of 0.005, and an R^2 of 98.2%. A comparison between the proposed MLP regressor model with several studies used the same dataset is illustrated in Table 10. The Table presents the MSE values obtained by different models, along with their corresponding

Table 9 Performance of the regression models using grid search method

| Models | MAE | MSE | MedAE | R^2 |
|---------------|--------------|--|---------------|--------------|
| KNN regressor | 0.009 | 0.0002 | 0.005 | 98.2% |
| DT regressor | 0.005 | 0.0001 | 0.0013 | 99% |
| SVR | 0.004 | 0.0001 | 0.0012 | 99.1% |
| MLP regressor | 0.003 | 2.8×10^{-5} | 0.0009 | 99.8% |

Table 10 Comparison between proposed MLP regressor model with several studies used the same dataset

| Studies | Model | MSE |
|------------------------|---|--|
| Ref [10] | ANFIS | 0.0029 |
| Ref [24] | NARNET | 0.1353 |
| Proposed MLP regressor | Parameters tuning for MLP regressor using grid search | 2.8×10^{-5} |

references. In [24], the NARNET model achieved an MSE of 0.1353, indicating its predictive performance in approximating the continuous-valued variable. The ANFIS model, on the other hand, achieved a substantially lower MSE of 0.0029, confirming its higher accuracy in predicting the target variable, according to [10]. The suggested MLP regressor model, however, outperformed both the NARNET and ANFIS models after parameter adjustment using the grid search approach. The suggested MLP regressor model has a low MSE of 2.8×10^{-5} , showing excellent precision in predicting the continuous-valued variable (highlighted in bold). The parameter tweaking procedure using grid search improved the model's accuracy even further, allowing it to outperform the other models assessed in the research. These MSE values give useful information about the models' performance, with the ANFIS model outperforming the NARNET model. However, the presented MLP regressor model, with its improved parameters, demonstrated excellent accuracy and attained the lowest MSE of all models tested. This highlights the efficacy of the proposed MLP regressor model, particularly when parameter tuning is applied using the grid search method, in accurately predicting the target variable and minimizing the prediction error.

From Table 10, the proposed MLP regressor model achieved better performance in the term of MSE than several previous studies.

Figures 12, 13, 14 and 15 illustrate the actual values vs. predicted values for KNN regressor model, DT regressor model, SVR model, and the proposed MLP regressor model, respectively, using grid search method. Visualizing the relationship between actual and predicted values in regression problems is an essential step for evaluating model performance and comprehending its behavior. This visualization yields invaluable insights, facilitating the assessment of prediction quality. Through this plot, this can effectively contrast the predicted values generated by the regression model with the actual values present in the dataset. This comparison swiftly reveals instances where the model's predictions align closely with actual observations and instances where discrepancies emerge. The plotted

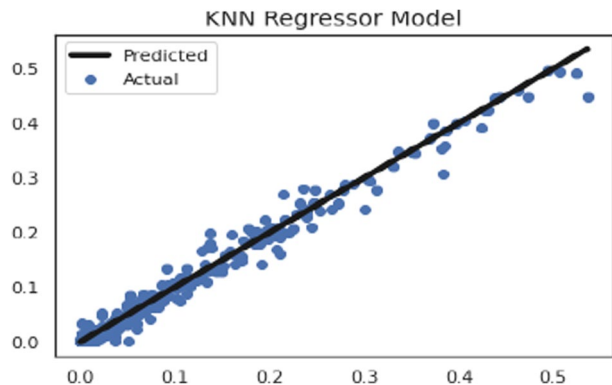
Fig. 12 Actual values vs predicted values for KNN regressor model

Fig. 13 Actual values vs predicted values for DT regressor model

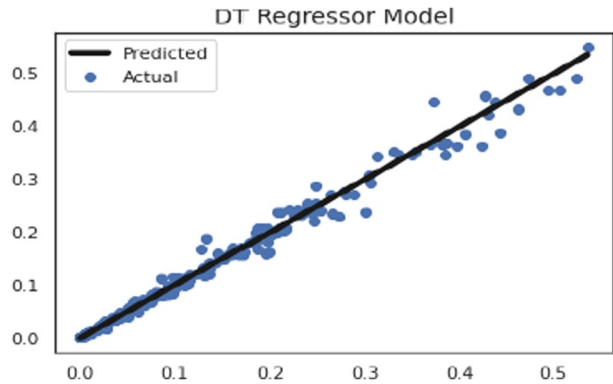


Fig. 14 Actual values vs predicted values for SVR model

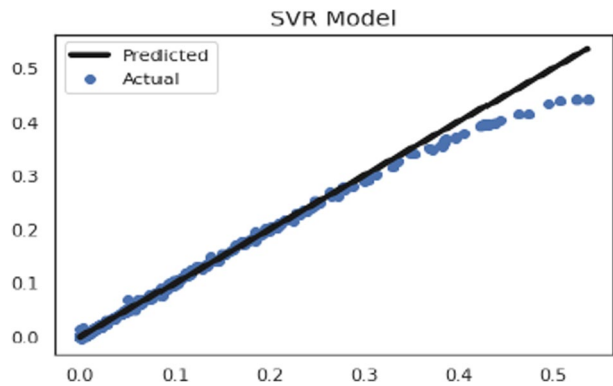
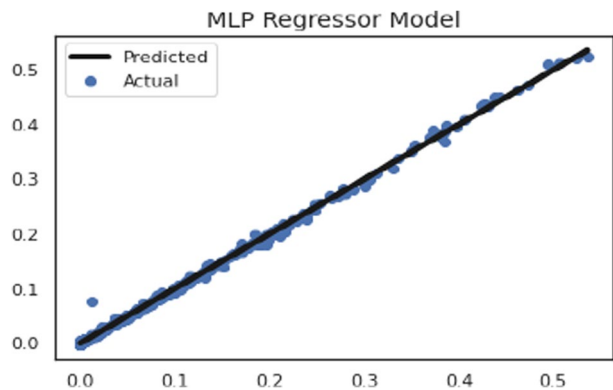


Fig. 15 Actual values vs predicted values for MLP regressor model



data points enable the identification of discernible trends or patterns governing the model's performance across distinct ranges of the target variable. Consequently, these visual cues shed light on the model's strengths and weaknesses, offering an opportunity to gauge its capacity to capture the underlying data relationships.

However, there are many potential limitations and challenges that should be considered. The specifics of the dataset used, and its representation require more detail on chemical

features and representation. Selection of other regions is required as well [19], considering the impact of climate change [40, 41]. In addition, the selection of models during the study may require prediction over a period of time, thus the use of LSTM and recurrent neural networks are mainly required [42, 43].

6 Conclusion and future work

In this paper, grid search method is used for tuning the parameters for four classification models and, for tuning the parameters for four regression models. The four classification models are RF, XGBoost, AdaBoost model, and GB model are used as classification models for predicting WQC. The four regression models are KNN regressor model, DT regressor model, SVR model, and MLP regressor model are used as regression models for predicting WQI. To assess the performance of the classification models, five assessment metrics were computed: accuracy, recall, precision, F1 score, and MCC. To assess the effectiveness of the regression models, four assessment metrics were computed: MAE, MedAE, MSE, and coefficient of determination (R^2). In terms of classification, the testing findings showed that the GB model utilizing the grid search approach produced the best results, with an accuracy of 99.5 percent when predicting WQC values. In regression, the experimental results illustrated that MLP regressor model using grid search method achieved the best results with R^2 equals 99.8% while predicting WQI values. In the future, we intended to use recurrent neural networks with LSTM to predict and the time serious analysis of the WQI and WQC in the presence of climate change variable.

Authors' contributions All authors are Equally Contributed.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Data availability Data is available at <https://www.kaggle.com/datasets/anbarivan/indian-water-quality-data>.

Code availability Available on Request.

Declarations

Conflicts of interest The authors declare that they have no conflicts of interest to report regarding the present study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.


References

1. Jain D, Shah S, Mehta H et al (2021) A Machine Learning Approach to Analyze Marine Life Sustainability. In: Proceedings of International Conference on Intelligent Computing, Information and Control Systems. Springer, pp 619–632
2. Clark RM, Hakim S, Ostfeld A (2011) Handbook of water and wastewater systems protection. In: Protecting Critical Infrastructure. Springer, pp 1–29. <https://doi.org/10.1007/978-1-4614-0189-6>
3. Hu Z, Zhang Y, Zhao Y et al (2019) A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture. *Sensors* 19:1420
4. Zhou J, Wang Y, Xiao F et al (2018) Water quality prediction method based on IGRA and LSTM. *Water* 10:1148
5. Waqas M, Tu S, Halim Z et al (2022) The role of artificial intelligence and machine learning in wireless networks security: principle, practice and challenges. *Artif Intell Rev* 55:5215–5261. <https://doi.org/10.1007/s10462-022-10143-2>
6. Halim Z, Waqar M, Tahir M (2020) A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email. *Knowl Based Syst* 208:106443. <https://doi.org/10.1016/j.knsys.2020.106443>
7. Wu J, Wang Z (2022) A Hybrid Model for Water Quality Prediction Based on an Artificial Neural Network, Wavelet Transform, and Long Short-Term Memory. *Water* 14:610
8. Lee S, Lee D (2018) Improved prediction of harmful algal blooms in four Major South Korea's Rivers using deep learning models. *Int J Environ Res Public Health* 15:1322
9. Liu P, Wang J, Sangaiah AK et al (2019) Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. *Sustainability* 11:2058
10. Hmoud Al-Adhaileh M, Waselallah Alsaade F (2021) Modelling and prediction of water quality by using artificial intelligence. *Sustainability* 13:4259
11. Bhardwaj D, Verma N (2017) Research paper on analysing impact of various parameters on water quality index. *Int J Adv Res Comput Sci* 8(5):2496–498
12. Malek NHA, Wan Yaacob WF, Md Nasir SA, Shaadan N (2022) Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. *Water* 14:1067
13. Slatnia A, Ladjal M, Ouali MA, Imed M (2022) Improving prediction and classification of water quality indices using hybrid machine learning algorithms with features selection analysis. In: Online International Symposium on Applied Mathematics and Engineering (ISAME22), vol 1. ISAME22, Istanbul-Turkey, pp 16–17
14. Deng T, Chau K-W, Duan H-F (2021) Machine learning based marine water quality prediction for coastal hydro-environment management. *J Environ Manage* 284:112051
15. Khullar S, Singh N (2022) Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation. *Environ Sci Pollut Res* 29:12875–12889
16. Abba SI, Pham QB, Saini G et al (2020) Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index. *Environ Sci Pollut Res* 27:41524–41539
17. Elbeltagi A, Pande CB, Kouadri S, Islam ARM (2022) Applications of various data-driven models for the prediction of groundwater quality index in the Akot basin, Maharashtra, India. *Environ Sci Pollut Res* 29:17591–17605
18. Asadollah SBHS, Sharafati A, Motta D, Yaseen ZM (2021) River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *J Environ Chem Eng* 9:104599
19. Nosair AM, Shams MY, AbouElmagd LM et al (2022) Predictive model for progressive salinization in a coastal aquifer using artificial intelligence and hydrogeochemical techniques: A case study of the Nile Delta aquifer, Egypt. *Environ Sci Pollut Res* 29:9318–9340
20. Garabaghi FH, Benzer S, Benzer R (2021) Performance evaluation of machine learning models with ensemble learning approach in classification of water quality indices based on different subset of features. *Res Square* 1:1–35. <https://doi.org/10.21203/rs.3.rs-876980/v2>
21. Hassan MM, Hassan MM, Akter L et al (2021) Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms. *Hum Centric Intell Syst* 1:86–97
22. Radhakrishnan N, Pillai AS (2020) Comparison of Water Quality Classification Models using Machine Learning. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES). IEEE, pp 1183–1188
23. Khan MSI, Islam N, Uddin J et al (2021) Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *J King Saud Univ – Comput Inform Sci* 34(8):4773–4781. <https://doi.org/10.1016/j.jksuci.2021.06.003>

24. Aldhyani THH, Al-Yaari M, Alkahtani H, Maashi M (2020) Water quality prediction using artificial intelligence algorithms. *Appl Bionics Biomech* 2020:1–12. <https://doi.org/10.1155/2020/6659314>
25. Khoi DN, Quan NT, Linh DQ et al (2022) Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam. *Water* 14:1552
26. Forests R, Breiman L (1999) Statistics Department University of California Berkeley. pp 1-29
27. Biau G (2012) Analysis of a random forests model. *J Mach Learn Res* 13:1063–1095
28. Wang S, Peng H, Liang S (2022) Prediction of estuarine water quality using interpretable machine learning approach. *J Hydrol* 605:127320
29. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining. pp 785–794
30. Prakash R, Tharun VP, Devi SR (2018) A comparative study of various classification techniques to determine water quality. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE, pp 1501–1506
31. Friedman JH (2002) Stochastic gradient boosting. *Comput Stat Data Anal* 38:367–378
32. Zhou Y, Mazzuchi TA, Sarkani S (2020) M-adaboost-a based ensemble system for network intrusion detection. *Expert Syst Appl* 162:113864
33. Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999) When is “nearest neighbor” meaningful? In: International conference on database theory. Springer, pp 217–235
34. Lu H, Ma X (2020) Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* 249:126169
35. Halim Z, Rehan M (2020) On identification of driving-induced stress using electroencephalogram signals: A framework based on wearable safety-critical scheme and machine learning. *Inf Fusion* 53:66–79. <https://doi.org/10.1016/j.inffus.2019.06.006>
36. Chen H, Huang JJ, McBean E (2020) Partitioning of daily evapotranspiration using a modified shuttleworth-wallace model, random Forest and support vector regression, for a cabbage farmland. *Agric Water Manag* 228:105923
37. Cheng Y, Peng J, Gu X et al (2020) An intelligent supplier evaluation model based on data-driven support vector regression in global supply chain. *Comput Ind Eng* 139:105834
38. Liao Z, Li Y, Xiong W et al (2020) An In-Depth Assessment of Water Resource Responses to Regional Development Policies Using Hydrological Variation Analysis and System Dynamics Modeling. *Sustainability* 12:5814
39. Tyagi S, Sharma B, Singh P, Dobhal R (2013) Water quality assessment in terms of water quality index. *Am J Water Resour* 1:34–38
40. Shams MY, Tarek Z, Elshewey AM et al (2023) A Machine Learning-Based Model for Predicting Temperature Under the Effects of Climate Change. In: Hassanien AE, Darwish A (eds) *The Power of Data: Driving Climate Change with Data Science and Artificial Intelligence Innovations*. Springer Nature Switzerland, Cham, pp 61–81
41. Elshewey AM, Shams MY, Elhady AM et al (2023) A Novel WD-SARIMAX Model for Temperature Forecasting Using Daily Delhi Climate Dataset. *Sustainability* 15:757. <https://doi.org/10.3390/su15010757>
42. Tarek Z, Shams MY, Elshewey AM et al (2023) Wind Power Prediction Based on Machine Learning and Deep Learning Models. *Comput Mater Contin* 74:715–732. <https://doi.org/10.32604/cmc.2023.032533>
43. Elshewey AM, Shams MY, Tarek Z et al (2023) Weight Prediction Using the Hybrid Stacked-LSTM Food Selection Model. *Comput Syst Sci Eng* 46:765–781. <https://doi.org/10.32604/csse.2023.034324>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Mahmoud Y. Shams¹  · **Ahmed M. Elshewey²** · **El-Sayed M. El-kenawy³** · **Abdelhameed Ibrahim⁴** · **Fatma M. Talaat^{1,5}** · **Zahraa Tarek⁶**

✉ Mahmoud Y. Shams
mahmoud.yasin@ai.kfs.edu.eg

Ahmed M. Elshewey
ahmed.elsheuey@fci.suezuni.edu.eg

El-Sayed M. El-kenawy
skenawy@ieee.org

Abdelhameed Ibrahim
afai79@mans.edu.eg

Fatma M. Talaat
fatma.nada@ai.kfs.edu.eg

Zahraa Tarek
zahraatarek@mans.edu.eg

¹ Faculty of Artificial Intelligence, Kafrelsheikh University, Kafrelsheikh 33516, Egypt

² Faculty of Computers and Information, Computer Science Department, Suez University, Suez, Egypt

³ Department of Communications and Electronics, Delta Higher Institute of Engineering and Technology, Mansoura 35111, Egypt

⁴ Computer Engineering and Control Systems Department, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt

⁵ Faculty of Computer Science & Engineering, New Mansoura University, Mansoura 35712, Egypt

⁶ Faculty of Computers and Information, Computer Science Department, Mansoura University, Mansoura 35561, Egypt