Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 4s (2024)

Computational Machine Learning Analytics for Prediction of Water Quality

Nitya Nand Jha[1], Rohit Kumar Singh[2], Sushila Sharma[3], Abhishek Kumar[4]

[1]Assistant Professor, Department of Civil Engineering, Rashtrakavi Ramdhari Singh Dinkar College of Engineering, Po- Ulao, Singhaul, Begusarai, Bihar, India, Pin- 851134

E-Mail id- nitya.n.jha@gmail.com

[2]Assistant Professor, Department of Civil Engineering,
Government Engineering College, Old Suta Mill Factory, Mairwa Road, Siwan, Bihar, India.Pin-841226.
E-mail id- rohitsingh372@gmail.com

[3]Assistant professor department of civil engineering Gaya College of Engineering, Gaya, Bihar, India Pin code- 823003

Email id- sushilasharma25@gmail.com

[4]Assistant Professor, Department of Mechanical Engineering, Rashtrakavi Ramdhari Singh Dinkar College of Engineering, Po- Ulao, Singhaul, Begusarai, Bihar, India, Pin- 851134

E Mail id- abhishekme10107@gmail.com

Article History:

Received: *15-04-2024*

Revised: *12-06-2024*

Accepted: *25-06-2024*

Abstract:

In terms of impacts on ecosystems, industry, people, and flora and fauna, water quality is paramount. Contamination and pollution have degraded water quality in recent decades. Predicting WQC and Water Quality Index (WQI) is the problem of this article; WQI is an important measure of water validity. This research use machine learning approaches to forecast WQI and WQC, and it does so by optimizing and tweaking the parameters of several machine learning models. Parameter optimization and tuning for four classification models and four regression models both make use of grid search, an essential tool in both contexts. To forecast WQC, classification models such as Random Forest (RF), Extreme Gradient Boosting (Xgboost), Gradient Boosting (GB), and Adaptive Boosting (Ada-Boost) are used. Predicting WQI is done using regression models such as K-nearest neighbour (KNN), decision tree (DT), support vector regression (SVR), and multi-layer perceptron (MLP). Data normalization and data imputation (mean imputation) were also executed as pretreatment steps to suit the data and make it convenient for any further processing. Seven characteristics and ninety-one cases make up the dataset used for this research. Five evaluation measures were calculated to evaluate the classification systems' effectiveness: accuracy, recall, precision, Matthews' Correlation Coefficient (MCC), and F1 score. A total of four evaluation metrics were calculated to measure the efficacy of the regression models: MAE, MedAE,

MSE, and R2. The results of the testing showed that the GB model yielded the most accurate predictions of WQC values (99.50%), making it the top performer in terms of categorization. The experimental findings show that the MLP regressor model got a value of 99.8 percent R2 when predicting WQI values, making it the best performing model in regression.

Keywords: Machine Learning, Water Quality, Prediction etc.

1. INTRODUCTION

To stay alive, water is an absolute must-have for all forms of life on Earth, including humans. Ensuring sufficient water quality is of the utmost importance for the survival of these creatures. Excessive pollution might have a devastating effect on aquatic animals' chances of survival and perhaps put their lives in jeopardy. A variety of natural water sources, such as rivers, lakes, and streams, may be evaluated according to specific quality requirements. To keep ecosystems in their natural state, water must meet certain criteria for various uses; for example, irrigation water must not be too salty and must not include pollutants that harm soil and plants. Furthermore, certain characteristics are necessary for industrial processes' water to fulfil their specific demands. Although surface and groundwater are naturally occurring and relatively affordable sources of freshwater, they are also susceptible to contamination from human and industrial activities. This is since it supplies data that is essential for managing and protecting predicting future water quality is a major challenge for environmental scientists. Quality prediction has been done using the use of physical principle-based simulations or empirical models, both of which may be resource-intensive and time-consuming. Conversely, with the advent of cutting-edge machine learning techniques, there has been a noticeable uptick in interest in the process of creating accurate and reliable models for water quality predictions. This study delves at the potential of machine learning algorithms to predict turbidity, pH, and dissolved oxygen concentration—indicators of water quality in various aquatic systems. We assess the relevant literature, review the relevant research, and then present case examples that show how well machine learning models predict water quality parameters. We also go over the limitations and downsides of using machine learning for this purpose. Machine learning may revolutionize water quality prediction, according to our findings, paving the way for better, more efficient use of water resources. This may be possible because to machine learning's revolutionary potential in water quality prediction.

2. LITERATURE REVIEW

Water quality index (WQI) and water quality class (WQC) prediction using machine learning algorithms is an expanding field of study. These algorithms are applied to a variety of water quality measures, including turbidity, total suspended solids (TSS), and dissolved oxygen (DO). The research of this topic is called "water quality prediction using machine learning." Scientists have trained and evaluated several algorithms using information sourced from a wide variety of locations and water sources. The research paper "Machine Learning-Based Ensemble Prediction of Water-Quality Variables" provides an example. "Using Feature-Level and Decision-Level Fusion with Proximal Remote Sensing" demonstrated the efficacy of machine learning regression methods and decision-level fusion in predicting water-quality attributes using data from three disparate Midwest bodies of water. This research was titled "Machine Learning-Based Ensemble Prediction of Water-Quality Variables." Similarly, the use of data from Norway's Brusdalsvatnet Lake in the paper "Emulating process-based water quality modeling in water source reservoirs using Machine Learning" demonstrates that the Long Short-Term Memory (LSTM) model, a subset of Machine Learning (ML), can effectively substitute process-based hydrodynamic and water quality models for use in water source management. These models were used to replicate the lake's water quality. The research study titled "Water Quality Prediction Using Machine Learning" used a dataset maintained

by the Central Pollution Control Board of India (CPCB) to evaluate the efficacy of several machine learning algorithms in predicting water quality. Based on data from the Directorate of Water Resources (DRE) of the State of Illizi, eight artificial intelligence algorithms were evaluated in the paper "Performance of machine learning methods in predicting water quality index based on the irregular data set: application on Illizi Region (Algerian Southeast)" to generate WQI predictions in the Illizi region, southeast Algeria. The data was supplied by the Directorate of Water Resources in the State of Illizi.

3. METHODOLOGIES

The following is a rundown of each process that goes into the production of our model:

Problem Identification: At this juncture, the objective is to find the issue statement. The task at hand is to use machine learning to forecast water quality.

Input: $x = [\sum_{i=0}^{n} I_n;$
$x_t rform = label.encoder(x);$
$y \leftarrow y_{train}, y_{test};$
$x \leftarrow x_{train}, x_{test};$
$n \leftarrow samples_o f_i mages;$
$TRUE_P \leftarrow TruePositive;$
$TRUE_N \leftarrow TrueNegative;$
$FALSE_P \leftarrow FalsePositive;$
$FALSE_N \leftarrow FalseNegative;$
$Features \leftarrow a, b;$
Accuracy: $\frac{TRUE_P+TRUE_N}{TRUE_P+TRUE_N+FALSE_P+FALSE_N};$
Precision: $\frac{TRUE_P}{TRUE_P+FALSE_P};$
Recall: $\frac{TRUE_P}{TRUE_P+FALSE_N};$
F1-Score : $\frac{2*TRUE_P}{2*TRUE_P+FALSE_P+FALSE_N};$
Activation : $\max[Accu, Prec, Reca, F1 - Sco, Sensi, Speci];$
while $yis \neq 0$ **do**
  if x_{test} *is Potable* **then**
    $accu \leftarrow \frac{TRUE_P+TRUE_N}{TRUE_P+TRUE_N+FALSE_P+FALSE_N};$
    $Preci \leftarrow \frac{TRUE_P}{TRUE_P+FALSE_P};$
    $reca \leftarrow \frac{TRUE_P}{TRUE_P+FALSE_N};$
    $f1 - sco \leftarrow \frac{2*TRUE_P}{2*TRUE_P+FALSE_P+FALSE_N};$
    $Sensi \leftarrow \frac{TRUE_P}{TRUE_P+FALSE_N};$
    $Speci \leftarrow \frac{FALSE_P}{FALSE_P+TRUE_N};$
else
  x_{test} is Not Potable
$accu \leftarrow \frac{TRUE_P+TRUE_N}{TRUE_P+TRUE_N+FALSE_P+FALSE_N};$
$Preci \leftarrow \frac{TRUE_P}{TRUE_P+FALSE_P};$
$reca \leftarrow \frac{TRUE_P}{TRUE_P+FALSE_N};$
$f1 - sco \leftarrow \frac{2*TRUE_P}{2*TRUE_P+FALSE_P+FALSE_N};$
$Sensi \leftarrow \frac{TRUE_P}{TRUE_P+FALSE_N};$
$Speci \leftarrow \frac{FALSE_P}{FALSE_P+TRUE_N};$

Fig 1. Algorithm for water quality classification

Data Extraction: To analyse, store, or process data at a separate location, data extraction involves gathering information from one or more sources. Our data was retrieved from the website in relation to the current condition.

Data Preprocessing: Improving the quality of data analysis is mostly dependent on processing the data. To generate valuable and applicable information, "data processing" involves collecting and transforming different parts of data.

Dealing with missing values: To fill in data that is lacking, there are a number of options. Most often, people will use means as a technique to deal with numerical columns that have missing values. However, means may not always be the best option when dealing with data that contains outliers. Therefore, the outliers must be addressed before the mean replacement method is used.

Water Quality Index (WQI):" "One comprehensive metric for water quality that takes all these aspects into account is the Water Quality Index (WQI). There have been nine different parameters used in the past to determine the WQI. When trying to determine the WQI in practice, Formula (1) is usually used.

$$wq1 = \sum_{i}^{n} \frac{= 1\ qi\ x\ wi}{\sum N = wi\ i}$$

Data Visualization: Data visualization refers to the act of presenting data visually to facilitate the discovery of trends, patterns, correlations, and other insights contained within the data (Fig. 2). matrix, we can use easily accessible characteristics to discover trends and create dependent features.

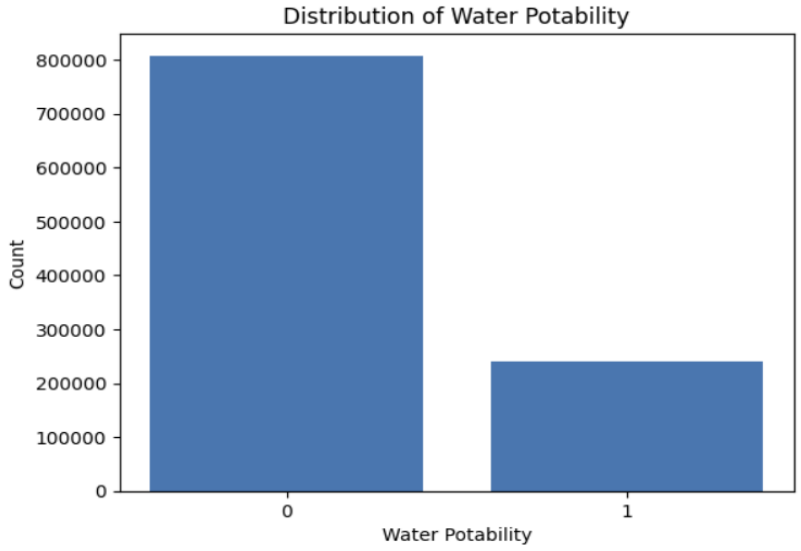

Fig 2 water potability

Correlation Analysis: It is possible to find the likely correlations between many factors with the use of a correlation matrix, which evaluates the correlation coefficients. You can see every possible value pairing in the table. Looked examined in the heatmap that the correlation created It is clear from looking at the study's Figure 3 that the correlation between all the characteristics is weak. Consequently, extracting any of the dataset's attributes is unnecessary.

4. DATASET

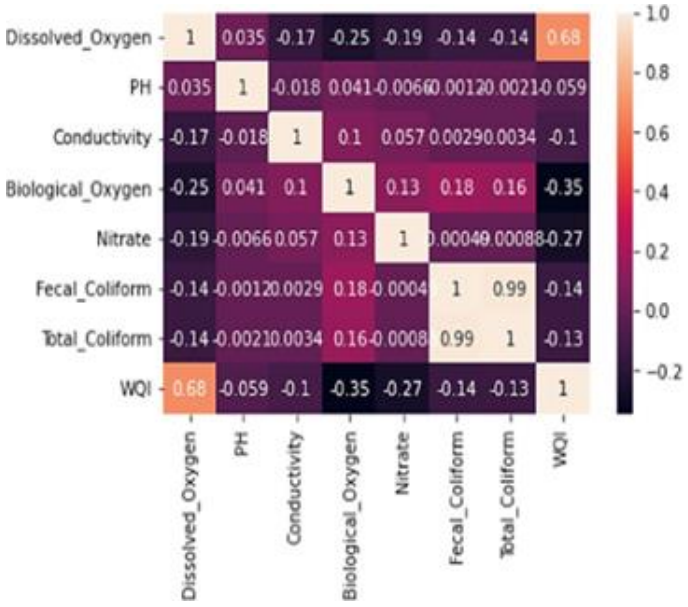

Fig 3 Heat map visualization of the feature correlations

The dataset used for this study is available at https://www.kaggle.com/datasets/anbar ivan/indian-water-quality-data. Several sites along rivers and lakes in India were surveyed between 2015 and 2022 for the dataset. To ensure the water is safe to drink, the Indian government gathered this data. In all, there are 2024 occurrences and 7 characteristics in the dataset. Dissolved oxygen, pH, conductivity, biological oxygen, nitrate, fecal coliform, and total coliform are included in the dataset. Characteristics of the data set include an indicator of the amount of oxygen dissolved in water, which is necessary for the survival of aquatic organisms, is dissolved oxygen. The pH scale indicates the degree of acidity or basicity of water by measuring its acidity or alkalinity. How well it of water, which provides data on the presence of dissolved solids and assesses its ability to conduct electrical current. An indicator of the level of organic pollution, the Biological Oxygen Demand (BOD) measures the amount of dissolved oxygen that microorganisms in water ingest. The Nitrate, which looks at the concentration of nitrate ions in water, which can be an indication of sewage or fertilizer contamination. Due to its reflection of the presence of coliform bacteria in the water, Fecal Coliform is a sign of faecal contamination. The sum of all coliform bacteria, whether they originate from feces or somewhere else, is known as total coliform. To ensure the dataset was usable and of high quality for the research, some preprocessing procedures were carried out. Both outliers and missing values are common in real-world datasets, and both procedures must be addressed. There is a lack of information on the data pretreatment phases in the provided context. Furthermore, statistical calculations on the dataset properties were also a part of the research (Table 2). In order to learn more about the distribution and characteristics of the data, these calculations may use metrics like the mean, standard deviation, minimum, maximum, and quartiles. In addition, as shown in Figure 5, the dataset's feature correlation matrix was examined. To find out whether there are any major connections or dependencies between the traits, the correlation matrix looks at how they relate to one another.

Table 1 Statistical calculation of the features

	Cnt	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

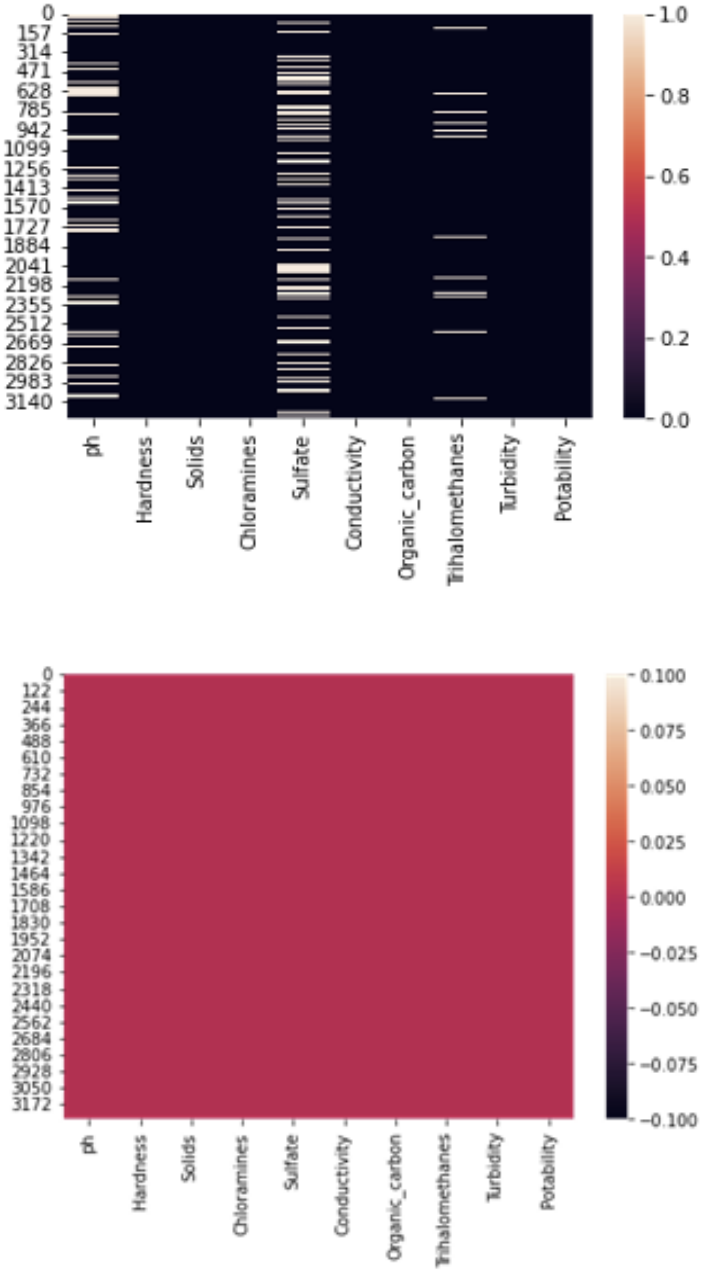

Fig 4 Heatmap before and after removing missing values.

The heatmap that may be found below (Fig. 4) displays the correlation that exists between the various characteristics.

Data Splitting: Separating the data into a training set and a testing set is an essential first step in evaluating the machine learning model's performance. To facilitate testing and training, the dataset was partitioned into two halves, with 33% of the data utilized for testing and 67% for training. For the model to make predictions or draw conclusions, it is necessary to establish a relationship between the dependent and independent factors. The results of the tests are then used to measure the machine learning algorithm's efficacy. Data partitioning allows it to compute accuracy metrics to evaluate the model's performance before applying it to real-world scenario simulations.

5. PREDICTION OF WATER POTABILITY USING ML ALGORITHMS.

Algorithm: The estimate of the water's potability was done using machine learning algorithms to attain this aim. We used algorithms to do both the classification and the regression. Several algorithms were used during our research.

Logistic Regression: The goal of this regression model is to use the values of the independent variables to estimate the probability of a certain outcome. while compared to linear regression, logical regression considers the logarithm of the outcome variable's probability while making predictions. When the dependent variable is continuous, linear regression is the method of choice for data analysis. The dependent variable may be described as a function of the independent variables by this transformation, but its range of values is restricted to just those between 0 and 1.

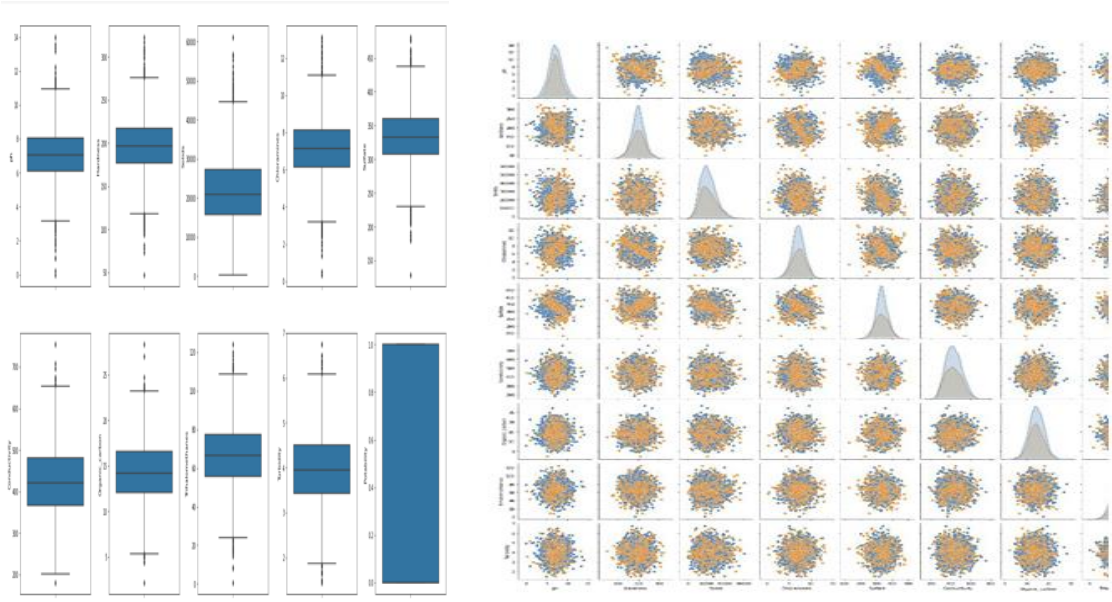

Fig 5 Visualizing data and checking for outliers.

As shown in (2), the sigmoid function is utilized in the process of doing analysis in logistic regression.

$$g\,(z) = \frac{1}{1 + e\text{^} - x}$$

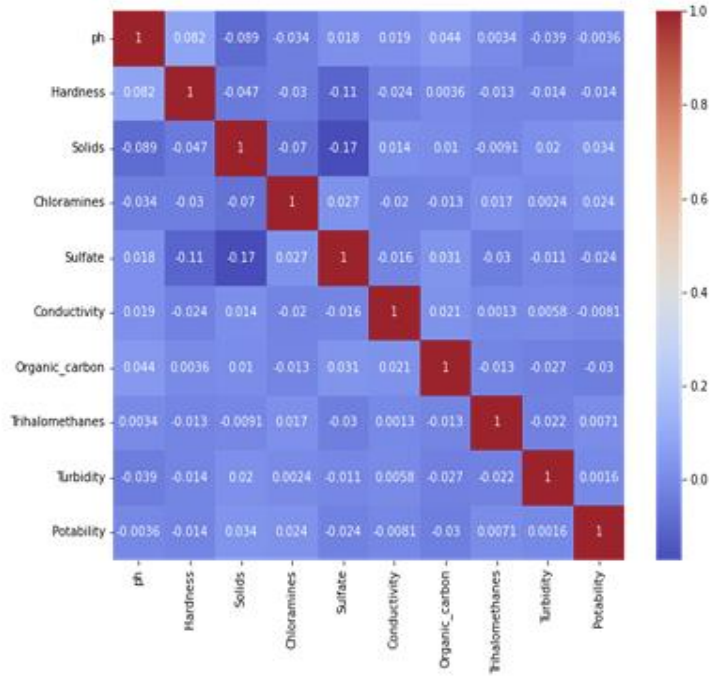

Fig 6 heat map

Support Vector Machine: Data classification, regression analysis, and outlier identification are all accomplished with the help of Support Vector Machines (SVM). In order to effectively partition the data into many groups, support vector machines (SVMs) are used. A hyperplane is a plane or route that maximizes the distance between two classes. How far away from the hyperplane each class's data points are from the hyperplane is what the margin is.

Decision Tree Classifier: Classification problems are often addressed by this kind of supervised learning technique, which is extensively used in machine learning. The choices and their related results are represented in a tree-like structure. Every node in the tree stands for a feature, and the values linked to that feature are grouped along each branch. The input instances are represented by the classes or categories that the tree's leaves represent.

Random Forest Classifier: Using a randomly selected sample of the training data and input attributes at each node, this classifier builds a succession of decision trees. This improves the tree's ability to forecast future data. The model's performance is enhanced in terms of its generalizability, thanks to the randomization that helps to minimize overfitting. The final forecast is either derived by averaging the forecasts of all the decision trees in the forest or by determining which predictions were most heavily voted. Each decision tree in the forest is trained independently.

XGBoost Classifier: The acronym XGBoost stands for "Extreme Gradient Boosting," the name of a multi-server distributed scalable machine learning algorithm. Specifically, it makes use of the GBDT algorithm. Some of the problems it can solve include regression, classification, and ranking. By far, the best machine learning framework out there.

AdaBoost Classifier: As part of an Ensemble Method, the boosting strategy called "Adaptive Boosting," or "AdaBoost," is used in machine learning. "Adaptive Boosting" is a possible acronym

for "AdaBoost." It is called "Adaptive Boosting" because it reassigns weights to each instance, giving more weights to instances that were incorrectly categorized. For the simple reason that doing so has the potential to increase the classification's overall accuracy.

K Neighbors: Based on the principle of supervised learning, K-Nearest Neighbor is among the most fundamental machine learning algorithms. Assuming the new instance or data is like existing cases is the operating assumption under which the K-NN technique functions. For each category, a new instance is created if it is most like an existing instance for that category. To properly categorize new data points, the k-NN algorithm must first recall all of the existing data and then use their similarity to previous data to determine how to do it. What this implies is that the K-NN method can readily sort newly acquired data into the correct suite category.

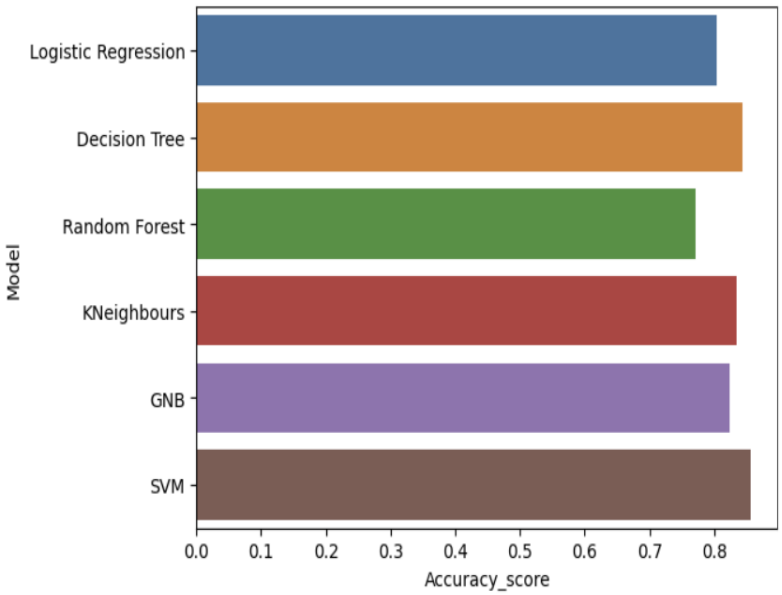

Fig 7 accuracy score

Measure: If you want to know what factors were considered while judging the model's efficacy, you may find them in the following list.

Precision: is the ratio of the number of contextual interpretations to the number of properly classified occurrences within a classifier. A measure of the level of accuracy associated with false alarms is FP, and TP, an abbreviation for "positive class," is calculated using Equation (3). Both ideas pertain to precision.

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

Accuracy: The most intuitive statistic is this one: it shows how many occurrences out of all the examples in the dataset have been properly classified, relative to the total number of examples. A dataset's accuracy may be determined by dividing its total occurrences by its true positive and true negative counts (which include all positive and negative values as well as any false positives or negatives) (Equation 4).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Recall: This metric has several names, such as sensitivity or the true positive rate. By calculating the true positive percentage, it finds out what fraction of the dataset's genuine positives are accurate positives. Equation 5 states that it may be calculated by dividing TP by the sum of TP and FN. Recall is useful when we want to find all positive examples with few false negatives.

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

F1 Score: When it comes to accessibility and precision, this is the sweet spot. It is a helpful statistic when both recall and accuracy should be considered since it finds a middle ground between the two. It is calculated as shown in Equation 6. F1 scores may be anything from 0 (very bad) to 1 (very good).

$$\text{F1 Score} = \frac{2\ x\ Precision\ x\ recall}{Precision\ x\ Recall}$$

Results for Algorithms: We used all the previously mentioned strategies to build the dataset-based regression and classification model. To evaluate the model, the hyperparameter tweaking method was used.

Table 2 comparison of different classifiers

	Model	Accuracy score
1	SVM	0.688540
2	XGBoost	0.670980
3	KNeighbours	0.653420
4	Decision Tree	0.645102
5	AdaBoost	0.634011
6	Logistic Regression	0.628466
7	Random Forest	0.628466

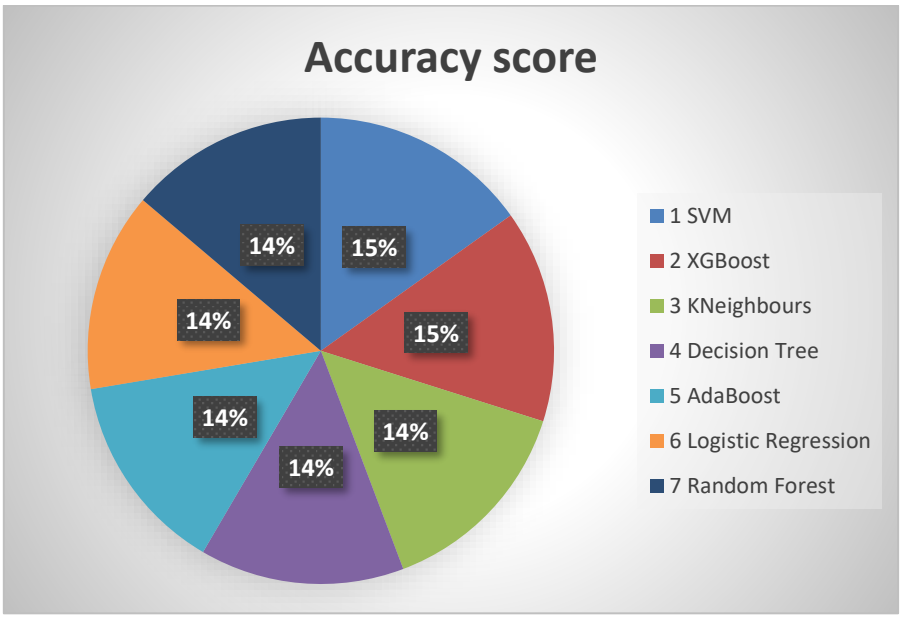

Fig 8 Accuracy score

Hyperparameter Tunning

When it comes to machine learning models, "hyperparameter tuning" refers to finding the sweet spot for each model's hyperparameters. Here, we talk about "hyperparameter tuning." The learning rate is a hyperparameter. The batch size, hidden layer count, and neuron count per hidden layer are some more examples. Because these model parameters cannot be learned during training, they must be supplied before training begins. Hyperparameter tweaking may be accomplished in several ways. Manual tuning, random search, grid search, and Bayesian optimization are a few examples of these strategies.

GridSearchCV: The user is asked to submit a grid of possible hyperparameters before GridSearchCV searches exhaustively over all possible

combinations of those hyperparameters. For every conceivable combination of hyperparameters, GridSearchCV runs a cross-validation test on the training data to assess the model's performance. The ideal hyperparameters are those that provide the greatest results in terms of performance.

RandomizedSearchCV: creates a set of hyperparameter combinations by randomly selecting from each of the distributions that are described for each hyperparameter. In order to evaluate the efficacy of a model, RandomizedSearchCV employs a technique known as cross-validation on the training data. This process is repeated for every conceivable combination. The hyperparameters that provide the greatest results are chosen as the best.

Bayesian optimization: Using probabilistic models, Bayesian optimization guides the search for the optimum hyperparameter combination, increasing the likelihood of achieving optimal outcomes. Because it is able to zero in on the most promising regions of the search space and find intriguing hyperparameters early on, this method outperforms grid search and random search. Both grid search and random search aim to cover the whole search space in one go.

Results of Hyperparameter Tunning: Classifiers like RF see an improvement in top-level accuracy and precision after hyperparameter tinkering, but a decline in accuracy in other cases, as seen in Table (4). In contrast, there are other contexts when precision degrades.

Table 3 Results of Hyperparameter Tuning

Model	**Accuracy before Hyperparameter tunning**	**Accuracy after Hyperparameter tunning**	
		Best Score	**Test Score**
SVC	0.688	0.605	0.628
xgboost	0.670	0.649	0.667
KNN	0.653	0.637	0.637
DT	0.645	0.632	0.63
Adaboost	0.634	0.637	0.64
Logestic Regression	0.628	0.605	0.6

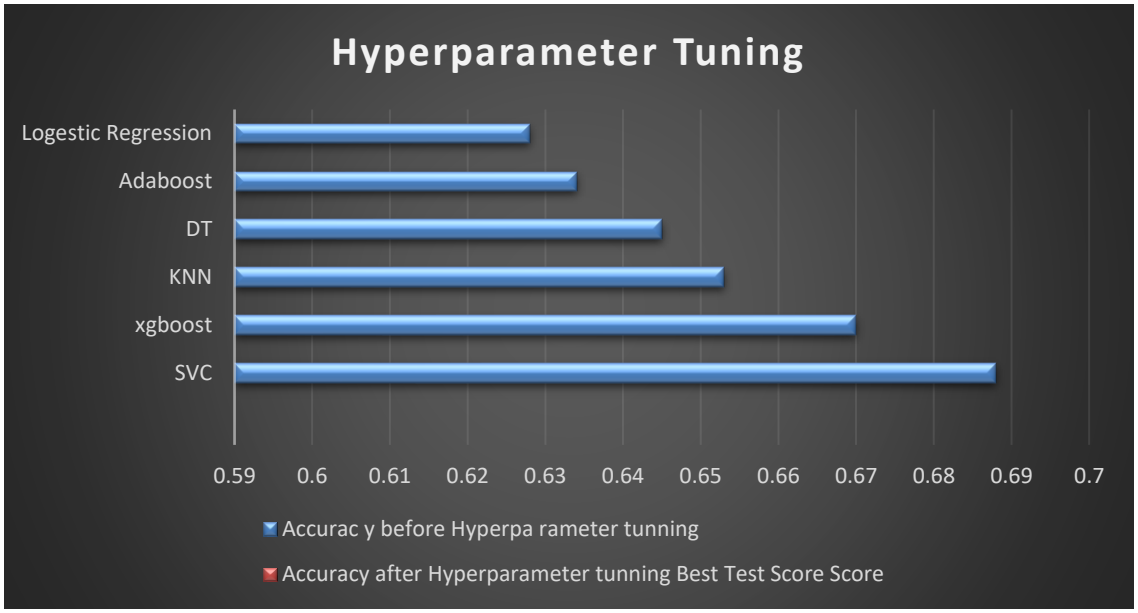

Fig 9 Hyperparameter Tuning

6. RESULTS

Using a dataset including information regarding water quality, this research tested, compared, and assessed the prediction capacities of five unique machine learning algorithms. This goal was accomplished by collecting values from popular datasets, including those for pH, hardness, solids, electrical conductivity (EC), and turbidity. Table 3 shows that the results showed that the employed models performed adequately in forecasting water quality data. On the other side, RF and XGB excel in terms of performance.

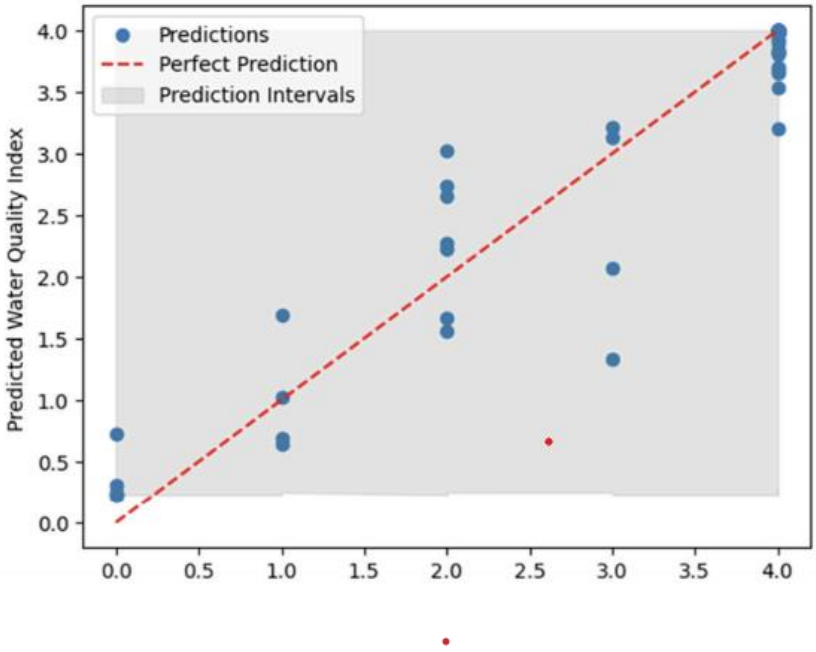

Fig 10 Actual water quality

Table 4 Classification report for different ML Algorithm

Model Name	Class label	Precision	Classification Report Recall	F1score	Accuracy
SVM	Not Potable	0.69	0.82	0.75	0.60
	Potable	0.55	0.37	0.44	
	Not Potable	0.68	0.89	0.77	0.64
XGBoost	Potable	0.61	0.31	0.41	
KNeighbours	Not	0.69	0.82	0.75	0.63
		0.55	0.37	0.44	
Decision Tree	Not Potable	0.66	0.90	0.76	0.63
	Potable	0.56	0.22	0.32	
AdaBoost	Not Potable	0.63	0.99	0.77	0.62
	Potable	0.62	0.04	0.07	
Logistic Regression	Not Potable	0.63	1.00	0.77	0.60
	Potable	0.00	0.00	0.00	
Random Forest	Not Potable	0.63	1.00	0.77	0.67
	Potable	0.00	0.00	0.00	

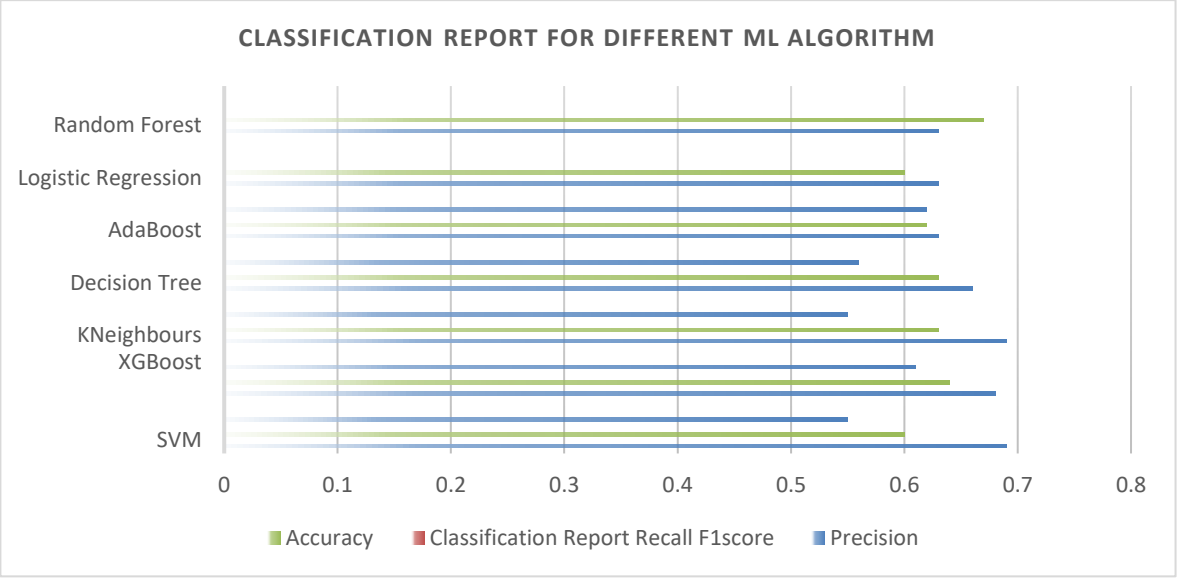

Fig 11 different ml algorithm

7. DISCUSSION

This research delves deeply into the difficulties and successes of coastal water quality prediction, focusing on how to use machine learning into parameter simulations for water quality. To preserve and protect coastal ecosystems, precise prediction models are crucial. These models must account for the effects of human activities such as urbanization, industrialization, coastal reclamation, and natural disasters including storms, floods, and erosion.

The paper accurately highlights how machine learning plays a crucial role in addressing numerical models' shortcomings. Problems with parameter choice, flexibility, and computing efficiency plague traditional models. One potential answer is the use of machine learning, which has been made possible by advancements in satellite remote sensing and UAV observation in recent times. Great

strides have been made in the processing of massive datasets and the extraction of relevant connections between satellite images and water quality measurements. The significance of remote sensing, especially satellite technology, is brought to light in the examination of techniques for collecting data on water quality. One example of an innovative technique to water quality data extraction is the use of spectral reference libraries and the investigation of optical properties. Optical feature-trained machine learning models outperform non-optical indicators when it comes to forecasting water quality metrics. Here, academics and practitioners may find a precise paradigm for inverting water quality data from remote sensing maps.

Additionally, this study offers a thorough synopsis of machine learning's uses in forecasting salinity, dissolved oxygen, chlorophyll-a, and other water quality metrics. We provide a comprehensive analysis of how well different machine learning techniques forecast these factors. Notably, it has been acknowledged that data properties and local circumstances dictate the selection of the most suitable algorithm. To help researchers find the best prediction models, we have included papers that compare various methods for certain parameters.

The characteristics of the simulated water quality may be used to pick from a variety of machine learning techniques. If you want to know where macro blooms are, you may use the Classification and Regression Tree technique [71]. Predicting levels of Escherichia coli and enterococci may be done using Decision Forest, Decision Jungle, and Boosted Decision Tree [47]. Using Extreme Gradient Boosting in conjunction with Long Short-Term Memory and Support Vector Regression (SVR) on a single model For the purpose of calculating Chl-a concentrations, Long Short-Term Memory works better. Predicting salinity concentrations may be done using Artificial Neural Networks, Gaussian Processes, and Support Vector Regression. When predicting the concentration of dissolved oxygen, Random Forest outperformed Support Vector Machine by a little margin. To forecast several water quality metrics, an ensemble machine learning model is an excellent option, as are Support Vector Regression, Extreme Gradient Boost, and Random Forest. The Random Forest method is able to pick important water quality indicators, and other models such as Extreme Gradient Boosting, Multi-layer Perceptron, Convolutional Neural Network, and Short-term Memory have also shown strong performance in forecasting WQI. A subset is better predicted by the Deep Neural Network technique.

One of the most important ways to measure water quality is via the water quality index (WQI), and this study attempts to clear up some of the confusion around the limitations of older WQI models. To improve WQI prediction, machine learning methods including Decision Tree, Random Forest, and Support Vector Machine have been used. While these strategies do provide desirable outcomes, we argue that they do not go far enough to improve WQI in and of themselves. We further the investigation of WQI prediction by talking about continuing attempts to reduce model uncertainty and enhance architecture.

A major emphasis of this work is the use of hydrodynamics into the process of water quality prediction. A comprehensive strategy is required to address the issues presented by nearshore waterways, which are affected by both coastal runoff and oceanic pressures. Storm surges, wave heights, and other dynamic parameters may be faithfully predicted using hydrodynamic prediction

models in conjunction with machine learning. By combining climate and risk data, our talk on the possibilities of Bayesian networks in coastal risk assessment offers a prospective view on coastal water quality prediction.

The study highlights how machine learning has revolutionized coastal water quality prediction. Our analysis of the shortcomings of existing models, the importance of varied datasets, and the implications of changing environmental circumstances suggests directions for further study.

8. CONCLUSION

This study summarises all the new developments in water quality prediction using machine learning. It is difficult to choose a single machine learning strategy with the highest performance, even after reviewing and comparing a large amount of literature. Machine learning models' performance might change greatly depending on the parameters and the area in question. An encouraging direction for future research is to conduct in-depth analyses of the properties of water quality parameters in an effort to develop more generally applicable methods. Using machine learning models that don't take physical and chemical processes into account makes it hard to generalize findings for complex coastal water quality predictions. Predicting processes incorporating critical elements outside the training dataset may be beyond the capabilities of models with regional features. To improve machine learning's prediction accuracy, we can do one of two things: (a) collect more data from more sources and increase its volume; (b) fill in missing data using interpolation methods like multiple-input denoising, k-nearest neighbors' input, or remote sensing satellite maps to invert optical characteristic parameters; and (c) conduct more research to quickly and effectively collect data on non-optical water quality parameters in coastal areas. Researchers, practitioners, and legislators engaged in environmental management and conservation will find this study to be an invaluable resource, as it provides an examination of the challenges and developments in coastal water quality prediction. To better protect the fragile coastal ecological balance, a new paradigm is emerging: numerical simulations that include machine learning. These models will be more precise and flexible.

Acknowledge

Authors acknowledge their parent institutions, [1] Rashtrakavi Ramdhari Singh Dinkar College of Engineering, Begusarai, [2] Government Engineering College, Siwan, [3] Gaya College of Engineering, Gaya for providing necessary resources and research support.

AUTHOR CONTRIBUTIONS

Nitya Nand Jha., Rohit Kumar Singh. and Sushila Sharma. performed the measurements, Abhishek Kumar was involved in planning and supervised the work. Nitya Nand Jha. and Rohit Kumar Singh. processed the experimental data, performed the analysis, drafted the manuscript and designed the figures. Nitya Nand Jha., and Rohit Kumar Singh. performed the calculations. Sushila Sharma.,. manufactured the samples and characterized them with spectroscopy performed the characterization and Abhishek Kumar implemented all analysis with help of machine learning. aided in interpreting the results and worked on the manuscript. All authors discussed the results and commented on the manuscript.

REFERENCES

[1] https://www.kaggle.com/datasets/adityakadiwal/ water-potabilit

[2] Patel, J., Amipara, C., Ahanger, T. A., Ladhva, K., Gupta, R. K., Alsaab, H. O., Althobaiti, Y. S., & Ratna, R. (2022). A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI. Computational Intelligence and Neuroscience, 2022, 1–15.

[3] Aldhyani, T. H. H., Al-Yaari, M., Alkahtani, H., & Maashi, M. S. (2020). Water Quality Prediction Using Artificial Intelligence Algorithms. Applied Bionics and Biomechanics, 2020, 1–12.

[4] Mohammed, H., Tornyeviadzi, H. M., & Seidu, R. (2022). Emulating process-based water quality modelling in water source reservoirs using machine learning. Journal of Hydrology, 609, 127675.

[5] Peterson, K., Sidike, P., Sidike, P., Hasenmueller, E. A., Sloan, J. M., & Knouft, J. H. (2019). Machine Learning-Based Ensemble

[6] Fu, Zhao, "Water Quality Prediction Based on Machine Learning Techniques" (2020). UNLV Prediction of Water-quality Variables Using Feature-level and Decision-level Fusion with Proximal Remote Sensing. Photogrammetric Engineering and Remote Sensing, 85(4), 269–280.

[7] Kouadri, S., Elbeltagi, A., Islam, A. R. M. T., & Kateb, S. (2021). Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). Applied WaterScience,11(12)

[8] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. Water, 11(11), 2210.

[9] Wang, R., Kim, J., & Li, M. (2021). Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. Science of the Total Environment, 761, 144057.

[10] Haghiabi, A. H., Nasrolahi, A., & Parsaie, A. (2018). Water quality prediction using machine learning methods. Water Quality Research Journal of Canada, 53(1),

[11] Jain D, Shah S, Mehta H et al (2021) A Machine Learning Approach to Analyze Marine Life Sustain ability. In: Proceedings of International Conference on Intelligent Computing, Information and Control Systems. Springer, pp 619–632

[12] Clark RM, Hakim S, Ostfeld A (2011) Handbook of water and wastewater systems protection. In: Pro tecting Critical Infrastructure. Springer, pp 1–29. https://doi.org/10.1007/978-1-4614-0189-6

[13] Hu Z, Zhang Y, Zhao Y et al (2019) A water quality prediction method based on the deep LSTM net work considering correlation in smart mariculture. Sensors 19:1420

[14] Zhou J, Wang Y, Xiao F et al (2018) Water quality prediction method based on IGRA and LSTM. Water 10:1148

[15] Waqas M, Tu S, Halim Z et al (2022) The role of artificial intelligence and machine learning in wire less networks security: principle, practice and challenges. Artif Intell Rev 55:5215–5261. https://doi. org/10.1007/s10462-022-10143-2

[16] Halim Z, Waqar M, Tahir M (2020) A machine learning-based investigation utilizing the in-text fea tures for the identification of dominant emotion in an email. Knowl Based Syst 208:106443. https:// doi.org/10.1016/j.knosys.2020.106443

[17] 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

[18] Lee S, Lee D (2018) Improved prediction of harmful algal blooms in four Major South Korea's Riv ers using deep learning models. Int J Environ Res Public Health 15:1322

[19] 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

[20] 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

[21] Malek NHA, Wan Yaacob WF, Md Nasir SA, Shaadan N (2022) Prediction of Water Quality Classi f ication of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. Water 14:1067

[22] 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

[23] 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

[24] Khullar S, Singh N (2022) Water quality assessment of a river using deep learning Bi-LSTM meth odology: forecasting and validation. Environ Sci Pollut Res 29:12875–12889

[25] Abba SI, Pham QB, Saini G et al (2020) Implementation of data intelligence models coupled with ensem ble machine learning for prediction of water quality index. Environ Sci Pollut Res 27:41524–41539

[26] 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

[27] 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

[28] Nosair AM, Shams MY, AbouElmagd LM et al (2022) Predictive model for progressive saliniza tion 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

[29] 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

[30] 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

[31] Radhakrishnan N, Pillai AS (2020) Comparison of Water Quality Classification Models using Machine Learning. In: 2020 5th International Conference on Communication and Electronics Sys tems (ICCES). IEEE, pp 1183–1188

[32] Khan MSI, Islam N, Uddin J et al (2021) Water quality prediction and classification based on principal component regression and gradient boosting classifier approach.

[33] 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

[34] 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 14:155

[35] 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

[36] Wang S, Peng H, Liang S (2022) Prediction of estuarine water quality using interpretable machine learning approach. J Hydrol 605:127320

[37] 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

[38] 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

[39] Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378

[40] Zhou Y, Mazzuchi TA, Sarkani S (2020) M-adaboost-a based ensemble system for network intru sion detection. Expert Syst Appl 162:113864

[41] Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999) When is “nearest neighbor” meaningful? In: International conference on database theory. Springer, pp 217–235

[42] Lu H, Ma X (2020) Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere 249:126169

[43] Halim Z, Rehan M (2020) On identification of driving-induced stress using electroencephalogram sig nals: 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

[44] Chen H, Huang JJ, McBean E (2020) Partitioning of daily evapotranspiration using a modified shut tleworth-wallace model, random Forest and support vector regression, for a cabbage farmland. Agric Water Manag 228:105923

[45] 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

[46] 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. Sus tainability 12:5814

[47] 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

[48] 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

[49] 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

[50] Elshewey AM, Shams MY, Tarek Z et al (2023) Weight Prediction Using the Hybrid Stacked-LSTM Food Selection.