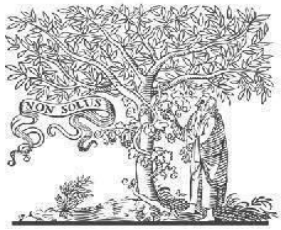
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

2024 IJIEMR. Personal use of this material is permitted. Permission from IJIEMR must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating newcollective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. No Reprint should be done to this paper; all copy right is authenticated to Paper Authors

IJIEMR Transactions, online available on 10th Dec 2024. Link

https://ijiemr.org/downloads.php?vol=Volume-13&issue= Issue12

DOI:10.48047/IJIEMR/V13/ISSUE12/02

Title: " ADVANCED WATER QUALITY CLASSIFICATION LEVERAGING SVM AND XGBOOST FOR OPTIMIZED PERFORMANCE IN ENVIRONMENTAL MONITORING"

Volume 13, ISSUE 12, Pages: 11- 23

Paper Authors
Boosa Divya, Dr.A.Pranayanath Reddy

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

ADVANCED WATER QUALITY CLASSIFICATION LEVERAGING SVM AND XGBOOST FOR OPTIMIZED PERFORMANCE IN ENVIRONMENTAL MONITORING

Boosa Divya[1], Dr.A.Pranayanath Reddy [2]

[1]PG Scholar, Department of CSE, Teegala Krishna Reddy Engineering College (Autonomous Institution), Medbowli, Meerpet, Saroornagar, Hyderabad–500097
[2] Associate Professor, Department of CSE, Teegala Krishna Reddy Engineering College (Autonomous Institution), Medbowli, Meerpet, Saroornagar, Hyderabad–500097.

ABSTRACT

One of the most valuable natural resources ever given to humans is water. The ecosystem and human health are directly impacted by the water quality. Water is used for many different things, including drinking, farming, and industrial uses. Over the years, numerous pollutants have put water quality in danger. Predicting and estimating water quality are now crucial to reducing water pollution as a result. Real-time monitoring is unsuccessful because conventionally, water quality is assessed using expensive laboratory and statistical processes. Low water quality calls for a more workable and economical solution. The proposed system builds a model that can forecast the water quality index and water quality class by utilizing the advantages of machine learning techniques. This proposed system is to develop a novel approach for water quality classification using Gradient Boosting Classifier. The method includes the calculation of the Water Quality Index, which is used as a measure of water quality. The proposed approach achieves a high Accuracy of 98%. The approach uses various water quality parameters and features such as PH, dissolved oxygen, temperature, and electrical conductivity to classify water into different categories. The model developed in this study is capable of predicting the water quality as Excellent, Good, Poor and Very Poor, which can be used for real-time monitoring and management of water quality. The results demonstrate the effectiveness and accuracy of the proposed approach in predicting water quality, high lighting the potential of machine learning techniques for water quality monitoring and management. The proposed approach can be used in various applications such as water treatment, environmental monitoring, and aquatic life management.

Keywords: Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees (GBT), and Support Vector Machine (SVM)

I. INTRODUCTION

The prediction of water quality through the application of machine learning techniques has recently emerged as a potentially fruitful and cutting-edge strategy for addressing the issues connected with the effective monitoring and management of water resources. Because of the ever-increasing need for pure and risk-free water, precise and effective forecasting of the criteria that determine water quality has

become an absolute necessity. Machine learning is a subset of artificial intelligence that offers powerful tools and techniques to analyze enormous volumes of data and make exact predictions. Because of this, it is a perfect option for improving the process of water quality evaluation and management

Traditional techniques of predicting water quality frequently rely on intricate mathematical models and arduous data processing, both of which can be time-consuming and resource-intensive. Traditional water quality predictions sometimes ignore crucial variables. Machine learning approaches are more efficient and automated. Machine learning algorithms need historical water quality data to find patterns and connections. pH, dissolved oxygen, turbidity, nutritional levels, and pollutants must be included. Because of this, they can accurately predict future water quality conditions.

This paper examines the enormous potential of machine learning for water quality forecasting. This cutting-edge technology helps us comprehend water quality dynamics in reservoirs, rivers, lakes, and municipal water sources. More precise water quality predictions allow proactive water resource management and contaminant mitigation to protect human health and the environment. This safeguards human and environmental health. Machine learning's ability to analyses massive amounts of data is one of its biggest benefits in water quality forecasting. Past methods have struggled to interpret and understand such big datasets, which can influence water quality. However, machine learning algorithms excel at finding meaningful patterns from large amounts of data, which helps us understand water quality trends and their causes.

This essay will examine water quality prediction machine learning algorithms. These include decision trees, random forests, support vector machines, neural networks, and ensemble techniques. By knowing the pros and downsides of each method, one may choose the best one for the situation and dataset. Machine learning offers potential but also challenges. Data quality and quantity are essential for powerful and accurate models. Incomplete or erroneous data may make accurate forecasts difficult and cast doubt on the inquiry. Thus, we shall emphasize data pretreatment and quality assurance in model construction.

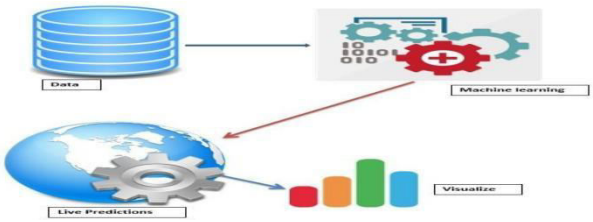


Figure1.1 Prediction of water in River

We will investigate the potential for machine learning to play an important part in the development of early warning systems for the contamination of water supplies. Machine learning algorithms can discover abnormalities and departures from typical conditions if the quality of water is continuously monitored in real time. This allows authorities and stakeholders to be notified of potential risks as soon as they arise. These kinds of preventative approaches have the potential to greatly lessen the impact of environmental dangers and water-borne diseases, the utilization of machine learning for the purpose of predicting water quality holds the potential to completely

transform how we manage and comprehend water resources. We will be able to improve water resource management, maintain aquatic ecosystems, and ensure the public's health if we take advantage of the power offered by this innovative technique. As we explore deeper into the complexities and possibilities of machine learning in water quality prediction, it is becoming increasingly clear that this technology has the potential to steer us in the direction of a more resilient and sustainable water future. We can pave the path for a society in which clean and safe water is accessible to all people while also maintaining our fragile water ecosystems if we continue our study and put our findings into practice.

The water that is stored in reservoirs is an essential resource that is necessary for maintaining ecosystems and keeping life going. These artificial or natural storage places play an essential part in giving a plethora of advantages to both people and the environment, and this function is critical to the success of the storage areas. In this section, we'll discuss some of the most important characteristics of reservoir water and its relevance, including the following: Freshwater, which is required for the continued existence of all living things, can be found in significant quantities in reservoirs. It is essential for the development, reproduction, and general.

Well-being of all living things, from microorganisms to plants and animals, that they have access to pure water, A broad variety of flora and wildlife can flourish in the varied aquatic ecosystems that are created by reservoirs. Aquatic life, such as fish, amphibians, and aquatic plants, flourishes in these conditions, which contributes to the general biodiversity and ecological balance of the environment, Consumption by humans is one of the most important applications for the water in a reservoir. Municipalities obtain water from reservoirs, treat it to make sure it is safe to drink, cook with, and clean with, and then distribute it to homes, schools, and businesses for a variety of uses, including those related to personal hygiene, cooking, and drinking, Agriculture can't function properly without water from reservoirs. It is utilized by farmers to irrigate their crops, which increases agricultural output and guarantees the availability of food. A sufficient water supply from reservoirs is necessary for the development of agriculture, which in turn helps communities to prosper.

play an important part in providing water for a variety of uses, which is necessary for the maintenance of life as well as ecosystems. It is imperative that these bodies of water be preserved and safeguarded if human health and the health of the environment are to be improved. Continuous monitoring of the water's quality assures that this invaluable resource will continue to be accessible and risk-free for future generations.

II. RELATED WORK

Predicting Arsenic (As) Exposure on Human Health for Better Management of Drinking Water Sources

Authors: -Minhaz Farid Ahmed, Chen Kim Lim, Mazlin Bin Mokhtar and Rd. Peri khanian Khirotdin**.**

Abstract: Chemical pollution in the trans boundary Langat River in Malaysia is common both from point and non-point sources. Therefore, the water treatment plants (WTPS) at the Langat River, which is toxic if

ingested via drinking water, is the subject of a pioneering study that predicts... of the main sources of drinking water to almost one-third of the population in Selangor state. Meanwhile, several studies have reported a high concentration of Arsenic (As) in The Langat River, which is toxic if ingested via drinking water, is the subject of a pioneering study that predicts... the as concentration in the Langat River based on time-series data from 2005–2014 To estimate the health risk associated with arsenic (As) ingestion via drinking water from the Langat River... Basin. Several time-series prediction models were tested and Gradient Boosted Tree (GBT) gained the best result. This GBT model also fits better to predict the as concentration until December 2024. The mean concentration of As in the Langat River for both 2014 and2024, as well as the carcinogenic and non-carcinogenic health risks of as ingestion via drinking water, were within the drinking water quality standards proposed by the World Health Organization and Ministry of Health Malaysia. However, the ingestion of trace amounts of as over a long period might be detrimental to human health because of its non- Biodegradable characteristics. Therefore, it is important to manage the drinking water sources to minimize arsenic (As) exposure risks.to human health.

Related work:

T Langat River Basin is one of the important river basins in the Selangor state of Malaysia covers approximately an area of 2409 km² within the latitudes 2°40'01.52"N to 3°16'01.50"N.and longitudes 101◦1902000 E to 102◦1 01000 E. In the upstream of Langat River Basin is the Titi Wangsa granite mountain range where the Langat River originates at the Gunang-Nuang and drains about 190 km via the Selangor and Negeri Sembilan states, as well as the Federal Territories of Putrajaya and Kuala Lumpur until it falls in the Strait of Malacca. In the Langat Basin, there are two dams, Semenyih and The Langat River, which is a potential reservoir of drinking water, apart from these dams, there are many ex-mining ponds all over the basin, especially at the Kuala Langat near the Paya Indah Wetlands. The topography of the basin is defined as both mountainous and flat, and the average elevation ranges from1440–400 m. The elevation of the central basin on average is below 200 m followed by less than 100 m at the lower basin. The rock beneath the Langat River Basin is igneous rock that is mainly granite. Therefore, the geology of the basin is defined as Hawthorn den Schist and Kenny Hill Formation (sandstone and phyllite). Tanah Coram (steepled), Rengan -Jerangan (urban land) and Tanah Gam but (peat soils) are the main soil types in the hilly upstream and flat midstream areas., and downstream, respectively. The average annual rainfall at the Langat River Basin is about 2483.62 ± 269.70 mm ranging from 2043.68 mm to 2832.40 mm during2005–2016.

Dataset: Monthly water quality data of arsenic (As) (µg/L) was collected from the four water quality monitoring sites. stations of the Department of Environment Malaysia (DOE) in the Langat River from January 2005 to December 2014. Accordingly, the monthly data of physicochemical parameters such as dissolved oxygen (DO, mg/L), pH, temperature (Temp, ◦C), salinity (SAL, ppt), and dissolved solids (DS, mg/L) were also obtained (2005–2014) from the same stations

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

of the Department of Environment Malaysia. The monthly missing values During2005–In 2014, yearly mean values of arsenic (As) and physico-chemical parameters were used as replacements. parameters data. The mean value of obtained As and physico-chemical parameters were found significant at a 95% confidence interval via the one-sample t-test (Table 2). SPSS software (Version21.0, IBM Corp., Armonk, NY, USA) was applied open form the descriptive statistics of as concentrations and physico-chemical parameters.

The Descriptive statistics include the calculation of the minimum (min.), maximum (max.), mean, and standard deviation. deviation (Std. Dev.) of the concentration of the water quality parameters in the Langat River. The standard deviation was calculated to observe the precision of each water quality parameter. Accordingly, Microsoft Excel 2016 (Microsoft Office Professional Plus2016, Microsoft Corporation, Redmond, Washington, USA) was used to produce the trend graphs of As and the physico-chemical parameters during 2005–2014 of the Langat River. Pearson's statistical correlation analysis was applied to estimate the correlations among the arsenic (As) concentrations and physico-chemical water quality parameters

One-sample test on monthly During 2005–2014, measurements were taken for as(µg/L) concentrations and physico-chemical parameters at the water sampling stations in Langat River.

Table 2.1physico-chemical parameters.

Water Stations	Parameters	N	t	df	Sig. (2-Tailed)	Mean Difference	95% Confidence Interval of the Difference	
							Lower	Upper
1L03	As (µg/L)	120	22.19	119	1×10^{-3}	3.64	3.31	3.96
	DO (mg/L)	120	89.06	119	1×10^{-13}	5.85	5.72	5.98
	pH	120	232.50	119	1×10^{-13}	6.98	6.92	7.04
	Temp °C	120	227.87	119	1×10^{-13}	28.52	28.27	28.77
	SAL (ppt)	120	2.13	119	0.04	0.12	0.01	0.24
	DS (mg/L)	120	19.02	119	1×10^{-13}	72.82	65.24	80.40
1L15	As (µg/L)	120	18.06	119	1×10^{-3}	4.74	4.22	5.26
	DO (mg/L)	120	71.15	119	1×10^{-13}	6.66	6.47	6.84
	pH	120	241.10	119	1×10^{-13}	7.05	6.99	7.11
	Temp °C	120	253.52	119	1×10^{-13}	27.15	26.94	27.36
	SAL (ppt)	120	23.70	119	1×10^{-13}	0.05	0.05	0.06
	DS (mg/L)	120	18.67	119	1×10^{-13}	57.29	51.22	63.37

Proposed algorithms:

Generalized Linear Model, Deep Learning, Decision tree, Random Forest, Gradient Boosted trees (GBT), and Support Vector Machine (SVM)

Methodology:

During the process of modelling to predict, RapidMiner Studio (Version 9.8, RapidMiner, NC., Boston, MA USA) was used. There are six models, namely Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees (GBT), and Support Vector Machine (SVM) that give the best top six results. Out of the six models, GBT has the best gain and best performance. The six steps of GBT processes are generalized as follows solely to run this set of data for QA (i.e., quality assurance) and QC (i.e., quality control)

purposes.

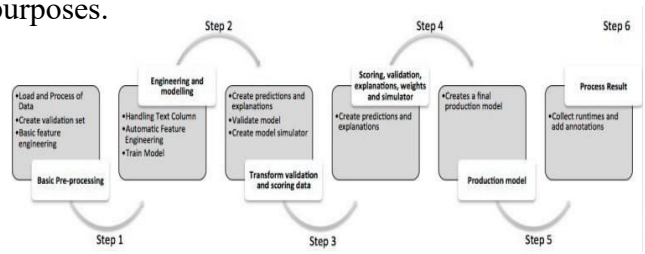


Figure2.1. Six steps of the Gradient Boosted

Trees (GBT)model's processes.

Friedman proposed the Gradient Boosted Trees (GBT) model in 2001. GBT produces competitive, highly robust, interpretable procedures for both regression and classification, especially appropriate for mining less than cleaning data. The gradient boosted tree (GBT) algorithm performed best on both the validation set and the challenge test set. For instance, the GBT algorithm, which outperformed... other models on both the validation set and the challenge test set, is also used for2017's Soccer Prediction Challenge [39], multi-solar power forecasting, and clinical mastitis based on milking data set. Figure 2. Six steps of the Gradient Boosted Trees (GBT) model's processes. Friedman proposed the Gradient Boosted Trees (GBT) model in 2001. GBT produces competitive, highly robust, interpretable procedures for both regression and classification, especially appropriate for mining less than cleaning data. The gradient boosted tree (GBT) algorithm performed best on both the validation set and the challenge test set. For instance, the GBT algorithm, which outperformed. other models on both the validation set and the challenge test set, is also used for2017's Soccer Prediction Challenge, multi-solar power forecasting, and clinical mastitis based on milking data set.

Forecasting Water Quality Index in Ground water Using Artificial Neural Network

Authors: -MonikaKulisz1, Justyna Kujawska, Bartosz Przysucha, and Wojciech Cel.

Abstract: Groundwater quality monitoring in the vicinity of drilling sites is crucial for the protection of water resources. Selected physicochemical parameters of waters were marked in the study. The water was collected from 19 wells located close to a shale gas extraction site. The water quality index was determined from the obtained parameters. A secondary objective of the study was to test the capacity of the artificial neural network (ANN) methods to model the water quality index in groundwater. The number of ANN input parameters was optimized and limited to seven, which was derived using a multiple regression model. Subsequently, Using the stepwise regression method, models with ever fewer variables were tested. The best parameters were obtained for a network with five input neurons (electrical conductivity, pH as well as calcium, magnesium, and sodium ions), In addition to five neurons in the hidden layer. The results showed that the use of the parameters is a convenient approach to modeling water quality index with satisfactory and appropriate accuracy. Artificial neural network methods exhibited the capacity to predict water quality index at the desirable level of accuracy (RMSE = 0.651258, R = 0.9992 and R2 = 0.9984). Neural network models can thus be used to directly predict the quality of groundwater, particularly in industrial areas. This proposed method, using advanced artificial intelligence, can aid in water treatment and management. The novelty of these studies is the use of the ANN network to forecast WQI groundwater in an area in eastern Poland that was not previously studied in Lublin.

Related work: Various studies examining the water quality index were presented. Hat zi kos et al. (2008) used neural networks with active neurons as a tool to predict the future (daily) indicators of seawater quality, such as water

temperature, pH, dissolved oxygen (DO), and turbidity. Experimental results yielded interesting information on the target variables and algorithm efficiency. Yesilnacar et al. (2008) predicted the nitrate concentration in groundwater using four parameters of temperature, electrical conductivity, groundwater level, and pH as input parameters in the ANN. The model tracked the experimental data very closely (R = 0.93). Hence, with the proposed model application, it is possible to manage groundwater resources in a more cost-effective and easier way. Hore proposed an ANN for predicting WQI. The input parameters of which were seasonal and positional variability of wastewater parameters under natural conditions. In turn, the ANN-based model developed by Salman et al. used monthly datasets on turbidity, total hardness, total dissolved solids, and electrical conductivity of the Shatt Al Arab River. The results obtained by Chen et al. indicate that ANN models exhibit very high efficiency in predicting the quality of groundwater, as well as the water in rivers, lakes, and wastewater treatment plants (WWTP).

Dataset: The analyzed water samples were taken from 17 dug wells located in the village of Syczyn and two deep water wells, which constituted drinking water intakes for the Wierzbica commune in the Lubelskie Voivodeship in Poland. The sampling area is flattened and its height is in the range of 178–181 m above sea level. The study covered a rural area where agricultural activities are carried out. The dug wells are located in the immediate vicinity of the Natural Gas Processing Plant, where the shale gas well is located.

The closest of the wells is 220 m and the farthest is 600 m away from the plant site. The two deepwater wells are located 4.9 km and 6.75 km away on the opposite sides of the gas drilling rig. The groundwater samples were stored in a laboratory refrigerator at 4 ◦C until subjected to the standard analysis procedures in accordance with the American Public Health Association. Table 1 specifies the methods and standards followed in the water quality parameters testing.

Table 2.2. Testing methods employed in the well water analysis

Parameter	Device	Norm
pH	ORION multimeter model: VERSA STAR (TermoFisher Scientific, London, United Kingdom; The Orion Navigator Pro computer software)	PN-EN 10390:1997
Electrical conductivity (EC)	ORION multimeter model: VERSA STAR (TermoFisher Scientific, London, United Kingdom; The Orion Navigator Pro computer software)	PN-EN 27888:1999
Total dissolved solids (TD) and total hardness (TH)	Titration kit	PN-C-04554-01:1971
Na, K, Ca and Mg concentrations	ICP-OES JY 238 Ultrace (Jobin Yvon-Horriba, France, Analyst software)	PN-EN ISO 17294-2:2016-11
Bicarbonate (IV) ions (HCO_3^-)	Titration kit	PN-C-04540-01:1990
Nitrates (V) (NO_3^-)	Foss FIAstar 5000 flow analyzer with photometric detection (Foss, Hillerod, Denmark, software SoFIA)	ISO 13395
Phosphates (V) (PO_4^{3-})	Foss FIAstar 5000 flow analyzer with photometric detection (Foss, Hillerod, Denmark, software SoFIA)	ISO/FDIS 15681-1
Cl concentration	Titration kit HACH 8051	Mohr's method Testing method
Sulfates concentration	Spectrophotometer (Hach Company; AMES, USA, Application Software Drinking Water DOC022.98.90372)	PBL/CH/28/06 edition 02 of 07.11.2011 based on HACH 8051

Proposed algorithms: Artificial Neural Network (ANN).

Methodology: The methodology for the development of the WQI, MLR and ANN models is shown in the flowchart

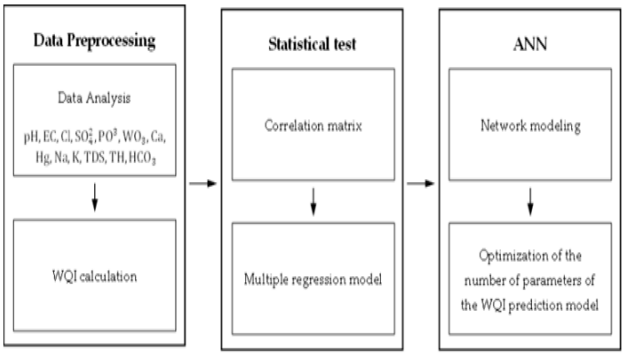


Figure2.2 flow chart models of MLR and ANN

Merits: The advantage of using the ANN model proposed in this paper is the easy assessment of groundwater pollution levels; in addition, it enables the avoidance of lengthy calculations involved in prevalent conventional WQI.

The use of neural network models is possible even in the case of data deficiency, e.g., due to the unavailability, difficulty or high-cost expense of obtaining the actual figures. It also allows for the reconstruction of missing data, reducing costs and saving time for otherwise indispensable analyses, which can be a strong economic factor in science

III. SYSTEM ANALYSIS

In the system we develop the Water Quality Classification Using Machine Learning with Gradient Boosting Classifier. The dataset used for this research was obtained from the Kaggle website, sourced from an Indian Government Website. The data is appropriate for the current research project since the characteristics required to construct the water quality index are available in this dataset. A water quality classification can be derived from the water quality index.

Before using the data for training, data need to go through data pre-processing, which refers to identifying and correcting errors in the dataset that may negatively impact a predictive model. The water quality index (WQI) was calculated using the dataset's most important parameters, which are Dissolved oxygen (DO), pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform. The WQI scores were then used to categorize the water samples. This research uses the weighted arithmetic water quality index method to calculate the WQI. The water quality has been classified into four groups. The next step is to train the Gradient Boosting Classifier model using the selected features and the calculated WQI. The model is trained using a portion of the water quality data, and the remaining data is used for testing.

The accuracy of the model is measured using various metrics such as Train Accuracy, Test Accuracy, Precision, Recall, and F1 Score. Water quality classification performance is evaluated using a confusion matrix. Since this study has four classes of water classification, a confusion matrix for multi-class classification is employed to portray the true classes of the data samples.

IV. IMPLEMENTATION

Gradient Boosting: Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

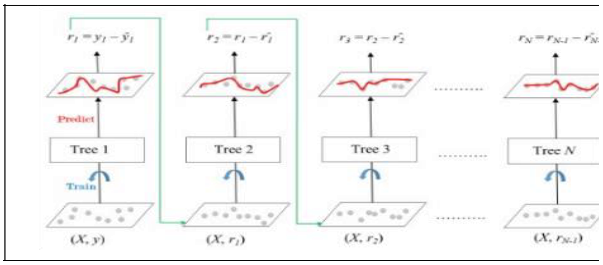


Figure: 3.1 ***Gradient Boosted Trees for Regression***

In contrast to Ada Boost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of the predecessor as labels. There is a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees). The below diagram explains how gradient-boosted trees are trained for regression problems The ensemble consists of M trees. Tree 1 is trained using the feature matrix X and the labels y. The predictions labeled y1(hat) are used to determine the training set residual errors r1. Tree 2 is then trained using the feature matrix X and the residual errors r1 of Tree 1 as labels. The predicted results r1(hat) are then used to determine the residual r2. The process is repeated until all the M trees forming the ensemble are trained. There is an important parameter used in this technique known as Shrinkage. Shrinkage refers to the fact that the prediction of each tree in the ensemble is shrunk after it is multiplied by the learning rate (eta), which ranges between 0 to 1. There is a trade-off between eta and the number of estimators; decreasing learning rate needs to be compensated with increasing estimators in order to reach certain model performance. Since all trees are trained now, predictions can be made. Each tree predicts a label and the final prediction is given by the formulary(pred) =y1 +(eta *r1) +(eta * r2) +……. + (eta * rN)

Algorithm:

Step 1: Let's assume X, and Y are the input and target having N samples. Our goal is to learn the function f(x) that maps the input features X to the target variables y. It is boosted trees i.e. the sum of trees. The loss function is the difference between the actual and the predicted variables

$$L(f)=\sum_{i=1}^{N} L(y_i, f(x_i))$$

Step 2: We want to minimize the loss function L(f) with respect of

$$\hat{f}_0(x) = \underset{f}{argmin}\ L(f) = \underset{f}{argmin} \sum_{i=1}^{N} L(y_i, f(x_i))$$

MmIf our gradient boosting algorithm is in M stages then to improve the f_m the algorithm can add some new estimator as h_m having $1 \leq \leq$

Step 3: Steepest Descent for M stage gradient boosting

$\rho_m - \rho_m g_m$ =The steepest Descent finds h_m where

Constant and known as step length and g_m is the gradient of loss function L(f)

$$g_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)}$$

Step4: Solution The gradient Similarly for M trees

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

$$f_m(x) = f_{m-1}(x) + \left(\underset{h_m \epsilon H}{argmin} \left[\sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) \right. \right.$$

$$\left. \left. \right) (x) \right.$$

the current solution will be $+ \; h_m(x_i))$

$$f_m = f_{m-1} - \rho_m g_m$$

V.RESULTS

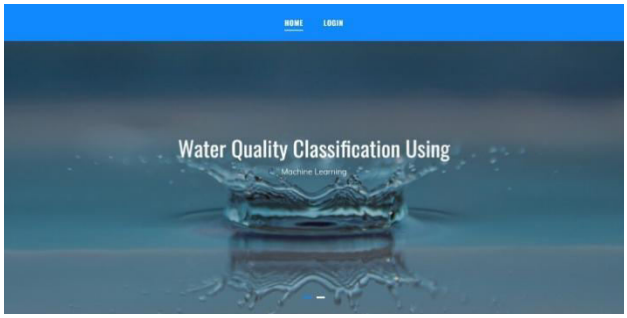


Figure: 1 Home Page

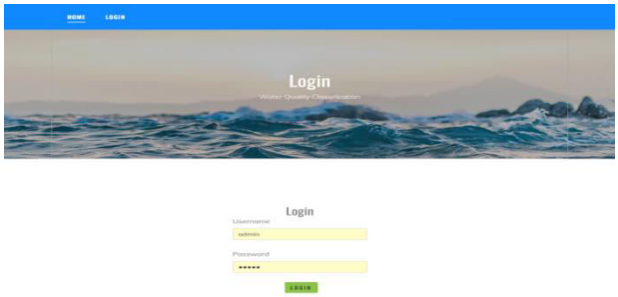


Figure: 2 Login Page

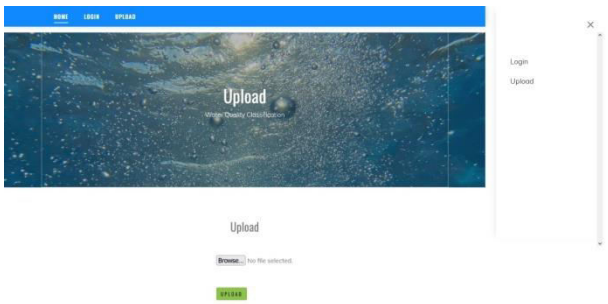


Figure:3 Upload Page

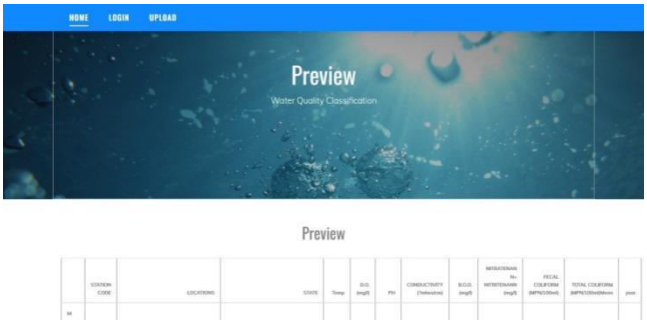


Figure:4 Preview Page

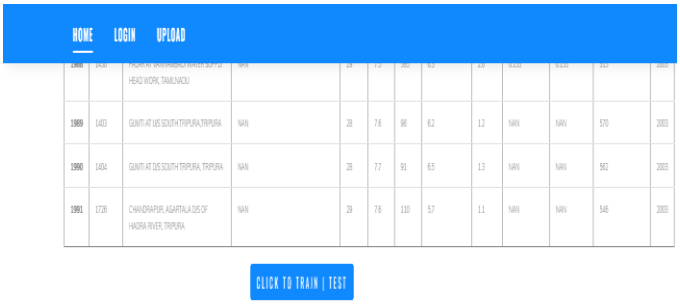


Figure:5 Train/Test On Preview Page

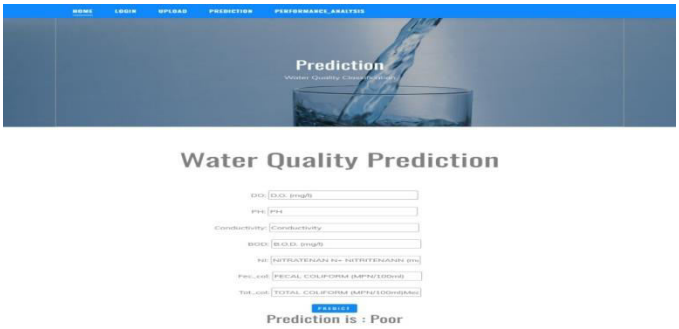


Figure:6 Prediction Test case-1

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Figure:7 Prediction Test case-2

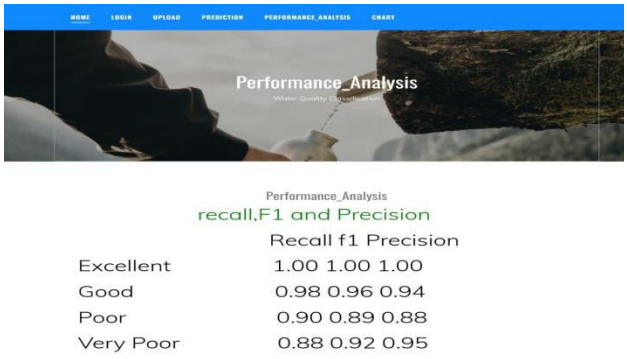


Figure:8 Performance Analysis Page

VI.CONCLUSION

Water quality is important in determining whether the water source is qualified for consumption. QI is essential to classify whether the water is safe for consumption. Rather than requiring expensive and complex analysis to test the water quality, this research uses Gradient Boosting Classifier, to predict water quality using readily available water quality parameters. The parameters employed for the classification algorithm are dissolved oxygen, pH, conductivity, biological oxygen demand, nitrate, fecal coliform, and total coliform. The outcome showed that Gradient Boosting Classifier outperformed the existing system even after the parameters had been tuned. In conclusion, this project highlights the significance of water quality and the need for an efficient and economical solution to monitor and manage it. The proposed approach, utilizing the advantages of machine learning techniques, provides an accurate and effective solution for predicting the water quality index and water quality class. The approach achieves a high Accuracy of 98%, indicating its potential for real-time monitoring and management of water quality. The model developed in This study can predict water quality as Excellent, Good, Poor, and Very Poor, enabling various applications such as water treatment, environmental monitoring, and aquatic life management. Overall, this project demonstrates the potential of machine learning techniques in the field of water quality monitoring and management, and it can be further improved and expanded to meet the increasing demand for efficient and reliable water quality management systems.

FUTURE ENHANCEMENT

There are several potential avenues for future work on this project. First, the proposed approach could be extended to include more water quality parameters and features to improve the accuracy and robustness of the model. Additionally, the model could be refined to incorporate more advanced machine learning algorithms to achieve even higher accuracy and to account for complex interactions between different water quality parameters.

Another potential area for future work is to integrate the proposed approach with real-time sensor data to develop a fully automated and continuous water quality monitoring system. This would involve deploying sensors at

various locations throughout a water system, collecting data, advanced machine learning algorithms to achieve even higher accuracy and to account for complex interactions between -time.

Moreover, it would be beneficial to explore the applicability of this approach to different types of water systems, such as lakes, rivers, and coastal waters, and to investigate the potential impact of climate change and human activities on water quality. Additionally, further research could be conducted to explore the economic feasibility and practical implementation of the proposed approach in various settings and applications.

Overall, there is a vast scope for future work on this project, and it has the potential to make a significant contribution to the field of water quality monitoring and management.

REFERENCES

[1] Illa Iza Suhana Shamsuddin, Zalinda Othman * and Nor Samsiah Sani *Water Quality Index Classification Based on Machine Learning: A Case from the Langat River Basin Model. Water2022, 14, 1552.

[2] Ahmed, M.F.; Lim, C.K.; Mokhtar, M.B.; Khirotdin, R.P.K. Predicting Arsenic (As) Exposure on Human Health for Better Management of Drinking Water Sources. Int. J. Environ. Res. PublicHealth.2021, 18, 7997.

[3] Kulisz, M.Kujawska, J.;Przysucha,B.;Cel, W. Forecasting water quality index in ground water using artificial neural network. Energies2021,14, 5875.

[4] Seyed Babak Haji SeyedAsadollah1, Ahmad Sharafati2,3,4*, Davide Motta 5, and Zaher Mundher Yaseen6. River Water Quality Index prediction and uncertainty analysis: a comparative study of machine learning models. Journal of Environmental Chemical Engineering.2020.

[5] Abba, S.I.; Hadi, S.J.; Sam men, S.S.; Salih, S.Q.; Abdulkadir, R.A.; Pham, Q.B.; Yaseen, Z.M. Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination. J. Hydrol.2020, 587, 124974.

[6] Ho, J.Y.; Afan, H.A.; El-Shafie, A.H.; Koting, S.; Mohd, N.S.; Jaafar, W.Z.; Sai, H.L.; AbdulMalek, M.; Ahmed, A.N.; Wan Mohtar, W.H.M.; et al. Towards a time and cost-effective approach to water quality index class prediction. Hydrol.2019,575, 148–165.

[7] Ahmed, U.; Mumtaz, R.; Anwar, H.; Shah, A.A.; Irfan, R.; García-Nieto, J. Efficient water quality prediction using super vised machine learning. Water 2019,11,2210.

[8] Chou, J.S.; Ho, C.C.; Hoang, H.S. Determining quality of water in reservoir using machine learning.Ecol.Inform.2018,44,57–75.

[9] Haghiabi, A.H.; Nasrolahi, A.H.; Parsaie, A. Water quality prediction using machine learning methods.WaterQual.Res.J.2018, 53, 3–12.

[10] M. Bouamar and M. Lad Jal, "Evaluation of the performances of ANN and SVM techniques used in water quality classification."

[11] F. Hassan baki Gara baghi, "Performance Evaluation of Machine Learning Models with Ensemble Learning approach in Classification of Water Quality Indices Based on Different Subset of Features," 2021, DOI: 10.21203/rs.3.rs 876980/v1.

[12] L. Li et al., "Interpretable tree-based ensemble model for predicting beach water quality," Water Research, vol. 211, Mar. 2022, DOI: 10.1016/j.watres.2022.118078. [13] N. Nasir et al., "Water quality classification using machine learning algorithms," Journal of Water Process Engineering, vol. 48, p. 102920, Aug. 2022, DOI: 10.1016/j.jwpe.2022.102920.

[14] D. Dezfooli, S. M. Hosseini-Moghari, K. Ebrahimi, and S. Araghinejad, "Classification of water quality status based on minimum quality parameters: application of machine learning techniques," Modeling Earth Systems and Environment, vol. 4, no. 1, pp. 311–324, Apr. 2018, DOI: 10.1007/s40808-017 04069.

[15] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, "Water Quality Prediction Using Artificial Intelligence Algorithms," Applied Bionics and Biomechanics, vol. 2020, 2020, DOI: 10.1155/2020/6659314.