International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Efficient ML-based Prediction Model for Analysis of Water Quality and Pollution in Yamuna River, India

Neetu Gupta[a], A. K.Sharma[b], Surendra Yadav[c]

Submitted: 28/07/2024 **Revised**: 07/08/2024 **Accepted**: 20/08/2024

Abstract— The level of pollution in the areas around the Yamuna River and industrial activity has significantly increased, thus it is crucial to assess the general state of the water quality in these areas. The production of agriculture, ecosystem services, and human health are all at risk due to the growing worldwide problem of water pollution. The unique characteristics of ensemble-based modeling and machine learning can provide a thorough understanding of the growing concerns about water quality. This study suggests a model that makes use of several methods, such as support vector machines, random forests, multinomial logistic regression, and extreme gradient boosting, to forecast pollution and water quality. It provides a thorough analysis of the water quality assessment and provides insights into the general state of the water quality in Delhi's Yamuna River, India, and near industrial areas. It achieves the best accuracy of 100% with the extreme gradient boosting technique**.**

Keywords— *MLP-AWQP, WQI, Yamuna River, LR, SVM, XGB, RF, water pollution*

I. INTRODUCTION

Water is one of the primary resources in our environment and society that must be kept pollution-free and healthy before its consumption. It becomes polluted and dirty due to industrial waste, sewage, pollution, etc. Such factors make water use difficult in normal conditions for different uses therefore, nowadays the quality of water requires a lot of attention for sustainability and purification. Advanced technologies such as Artificial Intelligence (AI) and Machine Learning (ML) have been utilized in numerous research studies for water quality assessment. These studies have utilized the Water Quality Index (WQI) as the primary parameter for evaluating the quality of rivers or lakes at various locations within their respective countries.

Many cities rely on the Yamuna River in Delhi, India for their water supply, but industrial discharges have severely contaminated it, endangering ecosystems, and human populations. With over 359 units producing untreated wastewater, industrial units in Delhi, Faridabad, Mathura, and Agra are a major source of pollution for the Yamuna River. The Yamuna River in Uttarakhand's Uttarkashi district. Many cities need to support industries, irrigation, and drinking water. Pollution is being addressed with the help of environmental laws, wastewater treatment, and public awareness campaigns. Sustainable usage of the Yamuna River necessitates efficient wastewater treatment, pollution control, and public participation. Water quality indices, which are vital for compiling data and directing pollution control actions, have been produced by many regions according to their specific demands.

This paper proposes the design of the ML-based Prediction Model for Assessment of Water Quality and Pollution (MLP-AWQP) model using Random Forest (RF), Multinomial Logistic Regression (MLR), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGB) techniques. It first collects the data and pre-processes it. The site locations are decided and then the MLP-AWQP system computes the WQI and sets the values ‘Good’, ‘Satisfactory’, and ‘Poor’ as labels. Furthermore, the model applies the extracted features to determine the correlation matrix and then performs the training and testing using four ML techniques. Lastly, the performance of the MLP-AWQP model is analyzed based on several factors.

This paper is architectured as follows: Section II presents a background that includes a brief on RF, LR, SVM, and XGB techniques and compares various existing water quality assessment methods and algorithms. Section III describes the system configuration and dataset used to implement the proposed MLP-AWQP model. Section IV explains the

Research Scholar, School of Engineering & Technology, Career Point University, Kota – 324005.India
‘School of Engineering & Technology, Career Point University, Kota – 324005, India
bDepartment of Computer Science and Application, Vivekananda Global University, Jaipur – 303012, India

MLP-AWQP model and its various stages. Section V provides a discussion of several experimental results and an analysis of the performance of the MLP-AWQP model. In Section VI, the paper concludes by discussing future work.

II. BACKGROUND

The literature review describes the RF, MLR, SVM, and XGB techniques. Furthermore, it provides a comparison among various existing water quality assessment techniques from year 2012 to 2024.

A. ML Techniques

This section provides an overview of ML-based techniques such as RF, MLR, SVM, and XGB techniques

1) RF technique: An RF is an ML algorithm and its popularity is due to its versatility and ease of use, as it can handle both regression and classification problems. The algorithm uses averaging to increase prediction accuracy and prevent overfitting. It achieves this by fitting several decision tree classifiers on different subsets of the dataset. The optimal split method is employed by forest trees. Fig. 1 depicts the working strategy of the RF technique how it takes the 'n' number of training samples to make the 'n' number of Decision Trees (DT) and how it applies voting with all 'n' DTs to make the prediction[21].

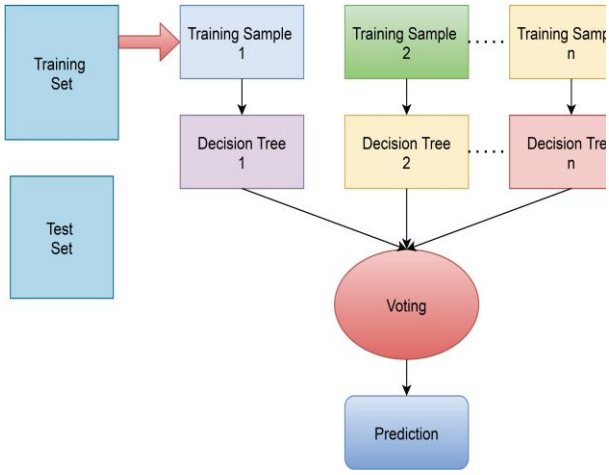

Fig. 1. Working procedure of RF technique

2) MLR technique: LR is a method for modeling the probability of a discrete result based on an input variable. Typically, LR models represent binary results, which means they can have only two values, such as true or false, yes or no, and so on. When estimating the likelihood of one of three or more alternative outcomes, MLR is utilized.

3) SVM technique: For tasks like regression, outlier identification, and linear or nonlinear classification, strong ML algorithms like SVM are used. To divide the classes in n-dimensional space, there may be more than one line or decision boundary. However, it needs to determine which decision boundary is best for classifying the data points using the hyperplane.

4) XGB technique: An ML method under ensemble learning is called XGB. For supervised learning tasks like regression and classification, XGBoost is a popular choice. XGBoost iteratively combines the predictions of multiple models, usually decision trees, to produce a predictive model..

B. Literature Review: Comparison among Various Existing Water Quality Assessment Methods

Several water quality assessment methodologies have been proposed to date using different types of AI and ML techniques. These methods are discussed and compared below. The design [1] applied ANN for computing the WQI and worked with a learning rate of 0.06A number of parameters were used in [2] to evaluate the WQI, including water temperature (WT), potential hydrogen (pH), calcium (Ca), electrical conductivity (EC), turbidity (T), dissolved oxygen (DO), total hardness (TH), nitrite (NO2), sulphate (S), phosphate (PO43-), magnesium (Mg), phosphorus (Po43-), sodium (Na), potassium (K), nitrate (NO3), total dissolved solids (TDS), total carbon (TC), biological oxygen demand (BOD) and chemical oxygen demand (COD).

In [3], the water quality was assessed using WQI, BOD, TDS, Concentration of Hydrogen Ions (CHI), DO, T, PO_4^{3-}, NO_3, Cl, WT, TH, EC, and ALKalinity (ALK). The study examined the effectiveness of ANN, SVM, and the Group Method of Data Handling (GMDH) for forecasting water quality components [4]. The results showed that the SVM was the most accurate model and that TANSIG and Radial Basis Function (RBF) as transfer and kernel functions performed better.

Adaptive Neuro Fuzzy Inference System (ANFIS), Weighted Average Ensemble (WAE), SVM, Auto Regressive Integrated Moving Average (ARIMA), Back Propagation Neural Network (BPNN), Neural Network Ensemble (NNE), Simple Average Ensemble (SAE), Multi-step ahead modelling techniques, and Adaptive Neuro Fuzzy Inference System (ANFIS) were adopted in another study [5].

The study [5] used features for the Hathnikund, Nizamuddin, and Udi stations, including DO, BOD, COD, Discharge (Q), pH, NH_3, and WT. When it came to performance accuracy, the ANFIS model outperformed the other three models, increasing average performance for Hathnikund and Nizamuddin stations by 7% and 19%, respectively, while SVM [6] outperformed the other models for Udi station, increasing the average performance by 16%.

A study on the efficiency of ML models for river water quality prediction was carried out, according to [7]. Numerous models have been applied to data analysis and prediction in numerous research. The main models in this category are wavelet-based techniques, fuzzy logic, artificial neural networks (ANN), SVM, hybrid neuro-fuzzy, hybrid ANN-ARIMA, genetic programming (GP), fuzzy logic, and hybrid neuro-fuzzy.

In another study [8], twelve hybrid Data Mining (DM) techniques were employed, including twelve-fold cross-validations, RF, M5P, Random Tree (RT), and Reduced Error Pruning Tree (REPT). It made use of parameters including WT, pH, Total Suspended Solids (TSS), NH_3, DO, BOD, and COD. The prediction of WQI was most affected by total solids and least by Fecal Coliform (FC), respectively. The next work [9] applied ML algorithms and WQI for water quality assessment using RF, K-Nearest Neighbor (KNN), and Tree-based Pipeline Optimization Tool (TPOT) and compared it with Automated ML.

In another study [10], Support Vector Regression (SVR), Multi Linear Regression (MLiR), Adaptive Neuro-Fuzzy Inference System (ANFIS), BPNN, and SVR were used to predict WQI. To increase the performance accuracy of the individual models, the NNE approach was used to propose the non-linear ensemble technique. The acquired results showed that the developed data intelligence models may reasonably be used to predict the WQI at the three stations, with the NNE outperforming other strategies with superior modeling outcomes. For Nizamuddin, Palla, and Udi (Chambal), the lowest values of root mean square ranged from 0.1213 to 0.4107, 0.003 to 0.0367, and 0.002 to 0.0272, respectively. Another method [11] [12] applied the SVM and fuzzy techniques for text extraction.

Extra Tree Regression (ETR) was established in a study [13] to predict monthly WQI values. SVR and DT Regression (DTR) were used to compare their performance. The prediction models were constructed using monthly input water quality data, which included BOD, COD, DO, EC, NO_3-Nitrogen (NO_3-N), NO_2-Nitrogen (NO_2-N), PO_4^{3-}, pH, WT, and T. More precise WQI forecasts were generated by the ETR model. The second-highest prediction accuracy was achieved with a set of input factors that included only BOD, T, and PO_4^{3-} content with $R_2^{test} = 0.97$ and $RMSE_{test} = 3.74$.

To examine the effectiveness of various models, including single and hybrid AI algorithms, for predicting river water quality, [14] offered a thorough review. The extreme learning machines, ANN, ANFIS, SVM, DT, and KNN were among these models. [15] used RF, NN, MLR, SVM, and BTM techniques for evaluating the water quality in different parts of India. Features including DO, Total Coliform (TC), BOD, NO_3, pH, and EC control the quality of water. It used a series of procedures, including feature correlation, applied classification, feature importance modeling, min-max normalization for data pre-processing, and RF for managing missing data.

According to the results [15], the main factors that went into the systematic classification of the water quality were NO_3, pH, EC, DO, TC, and BOD, with corresponding variable importance values of 74.7%, 36.8%, 81.4%, 105.7, 105.1, and 130.1. It used the datasets from Kadapa District's groundwater, India's rivers and lakes, the Pakistan Council of Research in Water Resources, Iranian Water Resources Management Company.

The study conducted [16] aimed to test the forecasting abilities of Gene Expression Programming (GEP), Artificial Neural Networks (ANN), and LiR approaches for simulating the monthly TDS and EC in the upper Indus River at two outlet stations. For both TDS and EC features, the correlation coefficient was more than 0.9. The GEP and ANN models continued to be the most dependable methods for EC and TDS prediction. The sensitivity analysis's findings showed that the input variables that affect TDS have an increasing trend: $HCO_3^- > Cl^- > Mg^{2+} > Na^+ > Ca^{2+} > SO_4^{2-} >$ pH. In contrast, the trend for EC was $HCO_3^- > SO_4^{2-} > Ca^{2+} > Cl^- > Na^+ >$ pH $> Mg^{2+}$.

The present methods for assessing water quality, especially the WQI, depend on expensive and time-consuming procedures for gathering data, and conventional predictive models find it difficult to adjust to changing environmental issues [17]. It applies sophisticated methods to promptly and precisely forecast WQI, which is essential for efficient water resource management. It applied the Latent Semantic Analysis (LSA) and XGB for prediction [17].

[18] provides a thorough analysis of the water quality assessment and provides insights into the general state of the water quality in Delhi's Yamuna River and near industrial areas. To do this, water samples from the Central Pollution Control Board (CPCB) were gathered throughout the course of the last eight years, from 2013 to 2021, and were then transformed into a format that could be read by machines to enable additional analysis. Numerous water quality indices, such as pH, DO, BOD, COD, nutrient levels, heavy metals, and other pertinent contaminants, were examined in these samples. To predict future trends and values of the water quality parameters, time series approaches are used. Table I explains the year-wise comparison among various existing methods based on country, river/lake, features, dataset used, ML technique applied, and performance.

C. *Limitations of the Existing Works*

- Need to apply with the other ML-based techniques [1]. Need to identify the sources of pollution and conduct ongoing monitoring of the lake's water and to prevent additional contamination of the greatest freshwater lake [2].
- Need to improve the WQI results [3]. Need to improve the accuracy [4] [8] [9] [16] [17].
- Extend with a combination of other algorithms and ensemble techniques and with DO and other WQI parameters [5].
- Need to improve the performance and accuracy of the proposed model by using different input parameters and advanced optimization algorithms. This can be done with an ensemble model consisting of several algorithms [10].
- To evaluate the WQI based on the World Health Organization (WHO) principles, different chemical and physical parameters must be used, thereby allowing the evaluation of more rivers with varying hydrological and climatic conditions [13].
- Necessary to combine the proposed method with other ML-based and deep-learning techniques [15].

III. Description of System Configuration and Dataset

The MLP-AWQP model was implemented on a system having the following configuration. It had the AMD Ryzen 5 3500U with Radeon Vega Mobile Gfx processor with 2.10 GHz, 8.00 GB RAM, and an x64-based 64-bit operating system.

The water quality dataset [19] was collected from the Delhi Pollution Control Committee (DPCC) in hard copy form for the years 2021, 2022, and 2023. The entire dataset was converted into the digital form, i. e. a .xls file. Water samples are gathered and analyzed for many factors from nine different sites along the Yamuna River, with an emphasis on industrial regions. The dataset had the water quality details of the Yamuna River, Delhi, India from three locations in Delhi, India such as ISBT bridge, Nizamuddin bridge, and Palla. It contains a total of 108 records for all three years and it had thirty-six records for each of three years. Furthermore, it had twelve records for each month of the year 2021, twelve records for each month of the year 2022, and twelve records for each month of the year 2023. Such bifurcation was maintained for each of the three locations in Delhi, India. This dataset is divided into a ratio of 70:30 for model training and testing, respectively.

IV. The Proposed Model Of ML-Based Prediction Model For Assessment Of Water Quality And Pollution

The proposed ML-based Prediction Model for Assessment of Water Quality and Pollution (ML-AWQP) is designed to evaluate the quality of the water by using RF, MLR, SVM, and XGB techniques. It performs a sequence of steps to determine the WQI for various samples of three years. Water samples are gathered and analyzed for several factors from nine different sites of the Yamuna River, India with an emphasis on industrial regions. After data pre-processing it extracts the features and determines the correlations among primary features for all three locations with month year and computes the label encoding as 'Good', 'Satisfactory', and 'Poor'.

TABLE I. Comparing Various Existing Methods for Water Quality Assessment

Ref No.	Year	Location	River / Lake	Features	Dataset	ML Technique (s) Used	Accuracy
[1]	2012	Malaysia	Kinta river	COD, pH, T, Mg, Ca, Cl, WT, EC, DO, BOD, K, MH_3-N, and WT. DS, Fe, RS, Na, SS, EC, NO_3-N, As, PO_4-P, TC, Zn, and Coli bacteria	Good Dataset	ANN	95.40%
[2]	2017	India	Loktak	Ca, Cl, F, S, TDS, Mg, PO_4^{3-}, Na, COD, K, NO_2, NO_3, TC, & BOD. 1.46 to 4.09 was the range of relative weight.	5 sampling sites	-	WQI ranges between 64% and 77%.
[3]	2017	Southern Iraq	Al-Gharraf	BOD, TDS, CHI, DO, T, PO_4^{3-}, NO_3, Cl, WT, TH, EC, ALK & WQI	5 sampling stations during 2015–2016. Monthly collection.	-	WQI ranges between 43% and 88.7%.
[4]	2018	Iran	Tireh river	Ca, Cl, EC, HCO_3, Mg, Na, SO_4, TDS, & pH	Good Dataset	ANN, GMDH, & SVM	33% to 98% for SVM with different features
[5]	2019	Delhi, India	Yamuna River	DO, BOD, COD, Q, pH, NH_3, & WT	3 stations data from CPCB	BPNN, ANFIS, SVM, ARIMA, SAE, WAE, & NNE	Increase in the average performance of NNE by 14%
[8]	2020	Iran	Talar River	COD, BOD, TSS, pH, WT, AN, DO, & WQI	Monthly data from two water quality monitoring stations spanning six years, from 2012 to 2018.	RF, M5P, RT, REPT & hybrid DM algorithms	R^2 = 0.941, RMSE = 2.71, MAE = 1.87, NSE = 0.941, PBIAS = 0.5
[9]	2020	-	River	pH, EC, TH, & WQI	Good Dataset	RF, KNN & TPOT	89% (RF), 68% (KNN), & 83% (TPOT)

[10]	2020	Delhi, India	Yamuna River	DO, pH, BOD, NH3, WT & WQI	CPCB	Ensemble of BPNN, ANFIS, SVR, & MLR	A noteworthy reduction in absolute error when compared to other models of 41%, 4%, and 3% for the Nizamuddin, Palla, and Udi (Chambal) stations, respectively.
[13]	2021	Hong Kong	Lam Tsuen River	PO_4^{3-}, pH, T, WQI, EC, BOD, COD, DO, and No_2-N	13 stations	SVR, ETR & DTR	R^2_{test} = 0.98, $RMSE_{test}$ = 2.99
[15]	2021	Kadapa, India, Pakistan & Iran	Rivers And Lakes	DO, TC, BOD, NO_3, pH, & EC	Several samples from different countries	ANN, RF, MLR, SVM, & BT	96.98% - 99.83%
[16]	2021	Indus River Basin (UIB)	Indus River	7 input parameters	360 TDS and EC monthly records for the last 30 years	GEP, ANN, & LiR	LiR: 0.90 (NSE) & 0.91 (R^2) for TDS, 0.94 (NSE) & 0.92 (R2) for EC.
[17]	2024	Delhi, India	Yamuna River	pH, DO, T, COD, BOD, WT, presence of pollutants, & WQI	CPCB dataset for the last 8 years	LSA, & XGB	95.20%
[18]	2024	India	Yamuna River	pH, DO, BOD, COD, & WQI	CPCB for the last 8 years	-	93.60%

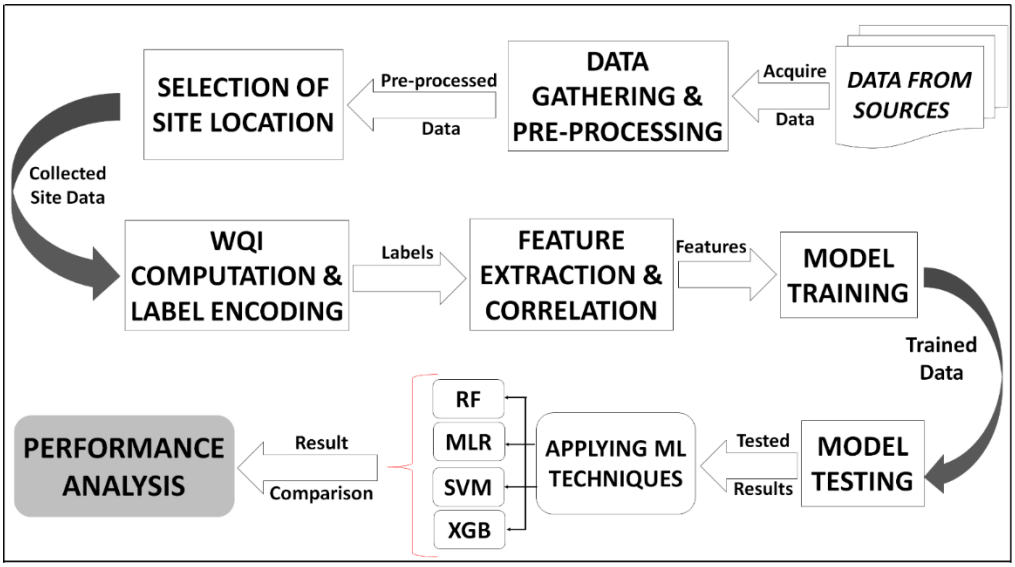

Fig. 2. The framework of the MLP-AWQP model

A. *Data Gathering and Pre-processing*

The first phase of the MLP-AWQP model collects the data [19]. It contains a total of five primary features such as pH, COD, BOD, DO, and FC. After data collection, it is pre-processed to remove the noise. It handles the incomplete and blank entries and prepares the data in its final form.

The pH scale indicates how basic or acidic water is. The degree of organic contamination in water bodies is represented by COD, a crucial metric for assessing water quality. BOD is the quantity of oxygen required for organic matter to break down. DO is the concentration of oxygen in water. A gram-negative, rod-shaped, facultatively anaerobic, non-sporulating bacterium is called a FC.

B. *Selection of Site Locations*

The MLP-AWQP model works with the dataset that was collected for Yamuna River, Delhi, India from three different locations such as ISBT Bridge, Nizamuddin Bridge, and Palla of Delhi.

C. *Computing WQI and Label Encoding*

The Yamuna River environment and public health are protected by pollution management measures that are implemented with the use of the WQI, a crucial numerical indicator that evaluates overall water quality conditions. A thorough understanding of water quality is crucial, making it a crucial component in assessing the condition of different water bodies to enhance management. The WQI is a crucial metric for effective water management. Multiple characteristics, including pH, DO, turbidity, COD, BOD, WT, and the presence of contaminants, must be considered when computing the WQI, which makes on-site data gathering necessary.

The WQI[20] is computed as given in equations (1), (2) and (3). The i^{th} parameter's sub-index is called Q_i, its unit weight is called W_i, there are n parameters total, its monitored value is called M_i, its ideal value is called I_i, and its standard value is called Si.

$$WQI = \sum_{i-1}^{n} \frac{WiQi}{\sum_{i-1}^{n} Wi} \quad (1)$$

$$Wi = \frac{K}{Si} \quad (2)$$

$$Qi = \frac{(Mi - Ii)}{(Si - Ii)} x100 \quad (3)$$

D. *Feature Extraction and Correlation*

After data pre-processing and site location selection, the MLP-AWQP model extracts the features from the dataset. This model considers all five primary features along with 'month' and 'year'. Table II lists all primary features along with their unit weights[20]. It is seen from Table II that the 'BOD' and 'FC' features have a minimum weight of approx. 0.95, and 'COD' and 'DO' features have a maximum weight of 1.90. The MLP-AWQP model determines the correlation among these features for each month of all three years and each of the three site locations so that it gets the result as 'Good', 'Satisfactory', or 'Poor'.

TABLE II. DEPICTING THE FEATURE NAMES ALONG WITH THEIR UNIT WEIGHTS

Feature Name	Unit Weight
PH	1.42
COD	1.90
BOD	0.95
DO	1.90
FC	0.9514

E. *Model Training*

The next step in the MLP-AWQP model is to train the RF, MLR, SVM, and XGB classifiers with the 70% training dataset of the water samples.

F. *Model Testing and Performance Analysis*

The trained MLP-AWQP model is further tested with the 30% testing datasets. The WQI is computed, and the water quality is assessed. Table III depicts the WQI rates along with their classification criteria as 'Good', 'Satisfactory', and 'Poor'. Furthermore, the performance of the MLP-AWQP model is analyzed and compared based on accuracy, recall, and precision.

TABLE III. DEPICTING THE WQI RATE ALONG WITH THEIR CLASSIFICATION CRITERIA

WQI Rate	Classification Criteria
0-50	Good
51-100	Satisfactory
More than 100	Poor

V. DISCUSSION ON EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

Various experiments have been performed on the MLP-AWQP model using the RF, MLR, SVM, and XGB techniques. The accuracy, recall, and precision were computed for all four ML techniques. Their results were compared and the XGB technique outperformed with 100% results for all three

performance metrics. All four ML techniques showed their association with the five primary features.

A. Performance Metrics

Equations (4), (5) and (6) depict the formulas of accuracy, recall, and precision, respectively.

$$Accuracy = Predicted\ Value / Actual\ Value \quad (4)$$

$$Recall = \frac{True\ Positive}{Predicted\ Results} \quad (5)$$

$$Precision = \frac{True\ Positive}{Actual\ Results} \quad (6)$$

B. ML Techniques and Feature Importance

Table IV depicts the importance of all five primary features for each of the four ML techniques. It is seen here that the pH and FC features are found as lowest and highest significant features with the RF technique. The DO and BOD features are found as the lowest and highest significant features with the MLR technique. The pH and COD features are found as the lowest and highest significant features with the SVM technique. Lastly, the pH and FC features are found as the lowest and highest significant features with the XGB technique.

TABLE IV. DEPICTING THE IMPORTANCE OF FEATURES WITH ML TECHNIQUES

ML Technique	**Dataset Features**				
	pH	**COD**	**BOD**	**DO**	**FC**
RF	0.055	0.112	0.118	0.15	0.447
MLR	-3.88	0.534	0.592	-0.42	0.232
SVM	-0.73	0.367	0.287	-0.72	0.055
XGB	0.00	0.089	0.159	0.025	0.696

C. Computation of Feature Correlation Matrix

Fig. 3 depicts the feature correlation matrix of the MLP-AWQP model in the range of [0, 1]. This matrix contains a total of 11 features including the five features from the dataset, the month and year of each data, three locations of Delhi, and label_Encoded features. The three locations of Delhi, ISBT Bridge, Nizamuddin Bridge, and Palla, are also considered here as the features in this correlation matrix.

The 'Label_Encoded' feature results in categorizing the quality of the water as either 'Good', 'Satisfactory', or 'Poor'. It can be seen from Fig. 3 that all the features show a correlation with each other. The water quality is checked for only one location at a time and the rest of the two locations become null. The same evaluation happened for the other two locations as well making the remaining two locations null in the same pattern. All the diagonal entries of the correlation matrix are 1 because each feature is entirely correlated with itself.

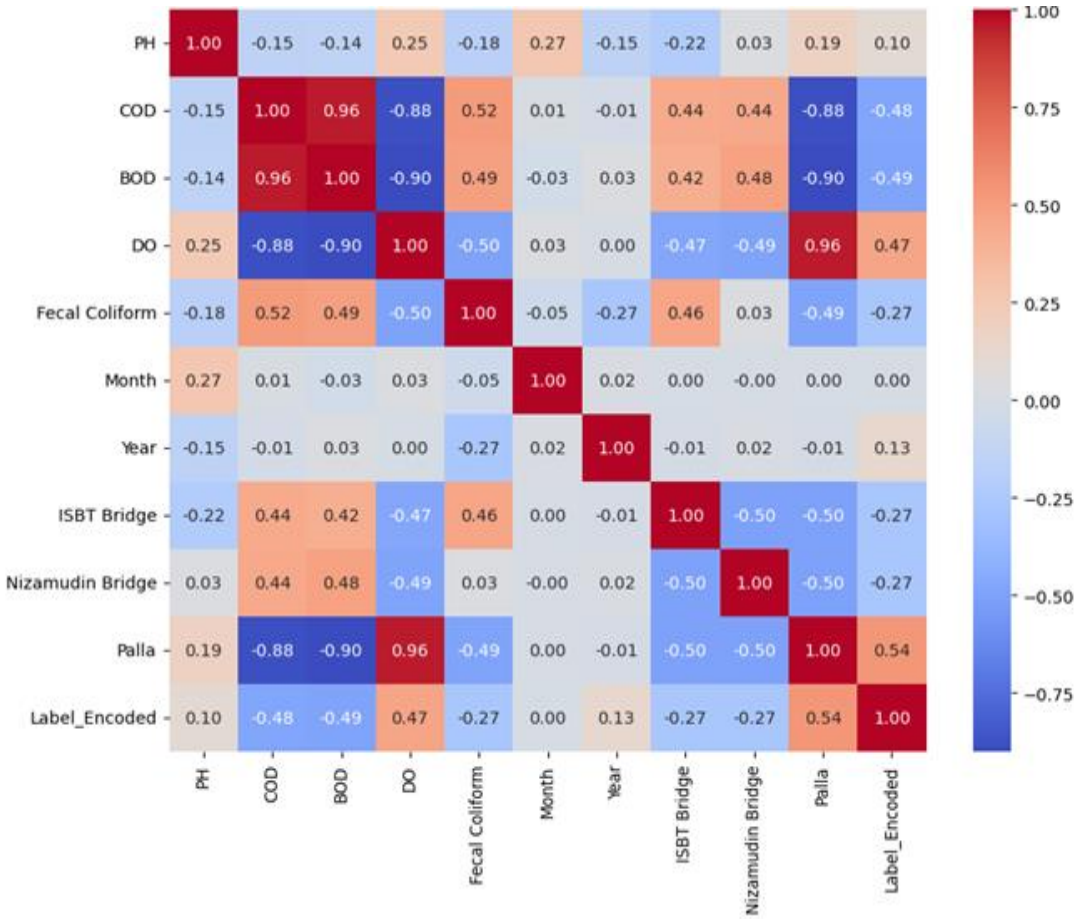

Fig. 3. Feature correlation matrix of the MLP-AWQP model

D. Performance Analysis with ML Techniques

Next, the performance of the MLP-AWQP is evaluated in terms of accuracy, recall, and precision using all four ML techniques. Table V depicts the comparison of their results. It is seen that the XGB technique outperforms and gives better results than RF, MLR, and SVM techniques. The maximum and minimum accuracy were achieved from MLR SVM, and XGB techniques, respectively. The minimum and maximum recall were achieved from MLR SVM, and XGB techniques, respectively. The minimum and maximum precision were achieved from MLR and XGB techniques, respectively. The worst results were achieved from the MLR technique for three performance metrics. Fig. 4 shows a 3-D graph depicting the comparison of performance metrics with the RF, MLR, SVM, and XGB techniques.

TABLE V. RESULTS OF WATER QUALITY WITH ML TECHNIQUES

ML Technique	**Accuracy (%)**	**Recall (%)**	**Precision (%)**
RF	95.45	95.45	97.72
MLR	90.90	90.90	82.68
SVM	90.90	90.90	91.77
XGB	100.0	100.0	100.0

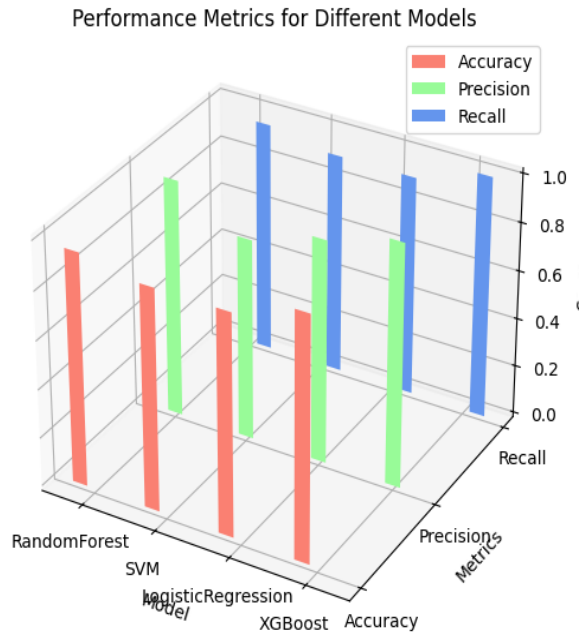

Fig. 4. Comparing the performance results with RF, MLR, SVM, and XGB techniques

The robustness of the proposed MLP-AWQP model and the generalized outcome were validated by WQI assessment standards. It results that these ML-based models are practical and affordable approaches to evaluating, managing, and forming policies related to Yamuna River's water quality.

VI. Conclusions and Future Work

The MLP-AWQP model that has been suggested explored the complex relationship that existed between environmental health, water quality, and the health of the Yamuna River ecosystem. The severe pollution levels, which are mostly caused by industrial discharges, emphasized how urgently strong action was required to protect this vital water source. Despite its importance, the WQI computation faces several obstacles, including laborious data collection procedures and rising expenses. Taking note of these constraints, the research set out on a ground-breaking expedition into the field of ML, utilizing its capabilities to transform WQI forecasts. In addition to streamlining the prediction process, the MLP-AWQP model achieved an accuracy of 100% using the XGB technique. This demonstrated how sophisticated predictive models can be used to address the changing complexity of water.

This work's primary contributions comprised three years of thorough data collecting and WQI estimations at important locations. This method not only improves our comprehension of water quality but also establishes a standard for further studies in the field. This research guides the face of environmental issues by shedding light on the way towards sustainable water management and the preservation of important water bodies such as the Yamuna River in India.

References

[1] N. M. Gazzaz, M. K. Yusoff, A. Z. Aris, H. Juahir, and M. F. Ramli, "Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors," Marine Pollution Bulletin, vol. 64, issue 11, 2012, pp. 2409-2420.

[2] R. D. Kangabam, S. D. Bhoominathan, S. Kanagaraj, and M. Govindaraju, "Development of a water quality index (WQI) for the Loktak Lake in India," Appl Water Sci, vol. 7, 2017, 2907–2918.

[3] S. H. Ewaid and S. A. Abed, "Water quality index for Al-Gharraf River, Southern Iraq," Egyptian Journal of Aquatic Research, vol. 43, issue 2, 2017, pp. 117-122.

[4] A. H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, "Water quality prediction using machine learning methods," Water Quality Research Journal, IWA Publishing, vol. 53, issue 1, 2018, pp. 3–13.

[5] G. Elkiran, V. Nourani, and S. I. Abba, "Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach," Journal of Hydrology, vol. 577, 2019, pp. 1-12.

[6] S. Puri and S. P. Singh, "A hybrid Hindi printed document classification system using SVM and fuzzy: an advancement," Journal of Information Technology Research, vol. 12, issue 4, 2019, pp. 107-131.

[7] T. Rajaee, S. Khani, and M. Ravansalar, "Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: a review," Chemometrics and Intelligent Laboratory Systems, vol. 200, 2020, pp. 1-25.

[8] D. T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, and N. Kazakis, "Improving prediction of water quality indices using novel hybrid machine-learning algorithms," Science of The Total Environment, vol. 721, 2020, pp. 1-15.

[9] K. Vijayaragavan, N. Praveen, M. V. Sudharsan, and P. S. Vijayan, "Machine learning model for water quality prediction using Python and AI

framework," International Journal of Advanced Research in Science, Communication and Technology, vol. 2, issue 3, 2022, pp. 360-365.

[10] S. I. Abba, Q. B. Pham, G. Saini, N. T. T. Linh, A. N. Ahmed, M. Mohajane, M. Khaledian, R. A. Abdulkadir, and Q.-V. Bach, "Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index," Environ Sci Pollut Res, vol. 27, 2020, pp. 41524–41539.

[11] S. Puri and S. P. Singh, "A fuzzy matching based image classification system for printed and handwritten text documents," Journal of Information Technology Research, vol. 13, issue 2, 2020, pp. 155-194.

[12] S. Puri and S. P. Singh, "Advanced applications on bilingual document analysis and processing systems," International Journal of Applied Metaheuristic Computing, vol. 11, issue 4, 2020, pp. 149-193.

[13] S. B. H. S. Asadollah, A. Sharafati, D. Motta, and Z. M. Yaseen, "River water quality index prediction and uncertainty analysis: A comparative study of machine learning models," Journal of Environmental Chemical Engineering, vol. 9, issue 1, 2021, pp. 1-15.

[14] N. Gupta and S. Yadav, "Review of various computational techniques for the assessment of water quality in the river," International Conference on Applications and Innovations in Intelligent Systems, Modelling and Optimization, VGU Campus, Jaipur, India, pp. 1-18, 2021.

[15] M. M. Hassan, L. Akter, M. M. Rahman, S. Zaman, K. M. Hasib, N. Jahan, R. N. Smrity, J. Farhana, M. Raihan, and S. Mollick, "Efficient prediction of water quality index (WQI) using machine learning algorithms," Human-Centric Intelligent Systems, Atlantis Press, vol. 1, issue 3-4, 2021, pp. 86-97.

[16] M. I. Shah, W. S. Alaloul, A. Alqahtani, A., Aldrees, M. A. Musarat, M. F. Javed, "Predictive modeling approach for surface water quality: development and comparison of machine learning models," Sustainability, vol. 13, issue 14, 2021, pp. 1-20.

[17] N. Gupta, S. Yadav, and N. Chaudhary, "Optimizing water quality prediction: a hybrid approach integrating latent semantic analysis and extreme gradient boosting for Yamuna River in Delhi, India," Power System Technology, vol. 48, 2024, pp. 1-19.

[18] N. Gupta, S. Yadav, and N. Chaudhary, "Time series analysis and forecasting of water quality parameters along Yamuna River in Delhi," International Conference on Machine Learning and Data Engineering, Procedia Computer Science, 2024, pp. 1-15.

[19] Dataset Link: https://www.dpcc.delhigovt.nic.in/

[20] T.H. Aldhyani, M. Al-Yaari, H. Alkahtani, M. Maashi, "Water quality prediction using artificial intelligence algorithms", Applied Bionics and Biomechanics 2020 (2020), 6659314.

[21] M.M. Hassan, Z.J. Peya, S. Mollick, M.A. Billah, M.M. Hasan Shakil,A.U. Dulla, Diabetes prediction in healthcare at early stage using machine learning approach, 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, Kharagpur, India, 2021,pp. 1–5.