Magna Scientia UCEVA
Vol. 4 N° 2 (2024) - P.P. 58 - 73
ISSN:2805-6884 e-ISSN:2805-6701
DOI:10.54502/msuceva.v4n2a6

Improving Water Quality Prediction in the Yamuna River, Delhi (India)

Perfeccionamiento de la predicción de la calidad del agua en el río Yamuna, Delhi (India)

Neetu Guptaa – Career Point University, Kota – India
Surendra Yadavb – University, Jaipur – India
Neha Chaudharyc – Manipal University, Jaipur – India

Open Access

Key words:

Yamuna River Water Quality Index (WQI), Hybrid approach, Latent Semantic Analysis (LSA), Extreme Gradient Boosting

Palabras clave:

Río Yamuna, índice de calidad del agua (WQI), enfoque híbrido, análisis semántico latente (LSA), Extreme Gradient Boosting.

Abstract

The Yamuna River, crucial for the water supply of several cities, faces a serious pollution problem due to industrial discharges, threatening both the health of ecosystems and the well-being of communities that depend on this resource. Current methods for assessing water quality, especially the quality index, are expensive and require considerable data collection time. In turn, traditional predictive models often fail to adapt to environmental changes, underlining the need for more advanced approaches that enable accurate and timely predictions of the water quality index, critical for effective water resource management.

In this research, machine learning techniques are used to make predictions on the water quality index, highlighting the limitations of existing models. The potential of various approaches is examined and an innovative hybrid methodology is proposed that combines Latent Semantic Analysis (LSA) for dimensionality reduction with Extreme Gradient Boosting, with the aim of improving the accuracy of predictions.

To conduct the study, water samples are collected from nine locations along the Yamuna River, focusing on industrial areas, and various parameters are analyzed. The calculated water quality index is then evaluated using various machine learning models as well as the proposed hybrid methodology. The evaluation criteria focus on accuracy, responsiveness, and the ability to predict the water quality index using limited but meaningful parameters.

The research results demonstrate the effectiveness of the hybrid methodology in predicting the water quality index, achieving a remarkable maximum accuracy of 95.2%, which is higher than other advanced models and techniques. This study provides valuable insights for water quality assessment, presenting an efficient and accurate data-driven approach essential for sustainable water resource management.

Resumen

El río Yamuna, crucial para el abastecimiento de agua de varias ciudades, se enfrenta a un grave problema de contaminación debido a los vertidos industriales, lo que amenaza tanto la salud de los ecosistemas como el bienestar de las comunidades que dependen de este recurso. Los métodos actuales para evaluar la calidad del agua, especialmente el índice de calidad, son costosos y requieren un considerable tiempo de recopilación de datos. A su vez, los modelos predictivos tradicionales a menudo no logran adaptarse a los cambios ambientales, lo que subraya la necesidad de enfoques más avanzados que permitan predecir de manera precisa y oportuna el índice de calidad del agua, fundamental para una gestión eficaz de los recursos hídricos.

En esta investigación, se utilizan técnicas de aprendizaje automático para realizar predicciones sobre el índice de calidad del agua, destacando las limitaciones de los modelos existentes. Se examina el potencial de diversos enfoques y se propone una metodología híbrida innovadora que combina el análisis semántico latente (LSA) para la reducción de la dimensionalidad con Extreme Gradient Boosting, con el objetivo de mejorar la precisión de las predicciones.

Para llevar a cabo el estudio, se recopilan muestras de agua en nueve ubicaciones a lo largo del río Yamuna, centrándose en áreas industriales, y se analizan diversos parámetros. Posteriormente, el índice de calidad del agua calculado se evalúa mediante varios modelos de aprendizaje automático, así como la metodología híbrida propuesta. Los criterios de evaluación se centran en la precisión, la capacidad de respuesta y la habilidad para prever el índice de calidad del agua utilizando parámetros limitados pero significativos.

Los resultados de la investigación evidencian la efectividad de la metodología híbrida en la predicción del índice de calidad del agua, alcanzando una notable precisión máxima del 95,2 %, superior a la de otros modelos y técnicas avanzadas. Este estudio proporciona valiosas perspectivas para la evaluación de la calidad del agua, presentando un enfoque basado en datos que resulta eficiente y preciso, esencial para la gestión sostenible de los recursos hídricos.

1. Introduction:

Water, as a vital component of our environment, plays a critical role in sustaining life and supporting various ecosystems. In an era marked by rapid urbanization, industrialization, and agricultural expansion, ensuring the availability of clean and safe water has become an imperative for global well-being. Proximity to rivers has been advantageous, providing water for various purposes. However, balancing river water use is crucial for sustainable resource management and protecting ecosystems. Pollution sources, including industrial discharges, agriculture, and sewage, vary by región [1, 3]. The River Yamuna faces severe pollution from industrial units in Delhi, Faridabad, Mathura, and Agra, with around 359 units releasing untreated wastewater. The Yamuna River in the Uttarkashi district of Uttarakhand. It's vital for several cities supporting drinking water, irrigation, and industries. Efforts are underway to address pollution through measures such as wastewater treatment, environmental regulations, and public awareness [4, 5]. Preserving the Yamuna River requires effective pollution control, wastewater treatment, and public involvement for sustainable use. Different regions have developed water quality indices tailored to their needs, essential for summarizing data and guiding pollution control measures.

The Water Quality Index (WQI) serves as a critical numerical index that assesses overall water quality conditions, to implement pollution control measures for safeguarding the Yamuna River ecosystem and human health. A crisp knowledge of water quality is essential, thus playing a pivotal role in evaluating

the state of various water bodies to improve their management. Computation of the Water Quality Index involves considering multiple parameters such as pH, dissolved oxygen, turbidity, chemical oxygen demand (COD), biochemical oxygen demand (BOD), temperature, and the presence of pollutants, necessitating on-site data collection. However, the earlier method of computing various parameters through samples was labor-intensive, and was associated with high financial costs [6]. Therefore, the WQI is indispensable for ensuring the repeated and effective monitoring of water body quality, especially in regions prone to frequent pollution. So, for early identification of such sources [7] and to predict WQI is the one of the majors concerned of the researcher in the past.

This work endeavors to explore several techniques to predict Water Quality Index, focussing on addressing the limitations of current models. Machine learning, a subfield of artificial intelligence, has demonstrated its efficacy in pattern recognition, data analysis, and prediction across diverse domains. By leveraging the capabilities of machine learning, this study aspires to enhance the accuracy and timeliness of WQI predictions, contributing to more effective water resource management strategies [8].

This research aims to evaluate quality of water in the surroundings of industrial areas along the Yamuna River in Delhi, employing various parameters. The collected data is then utilized to compute WQI by employing numerous models namely, logistic regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes (NB), and XGBoost. To enhance the results, a novel hybrid methodology is proposed, integrating Latent Semantic Analysis and Extreme Gradient Boosting. Latent Semantic Analysis performs dimensionality reduction on the dataset features through singular value decomposition, enhancing feature representation. The improved features are then fed into the Extreme Gradient Boosting technique for further prediction. Extreme Gradient Boosting (XGBoost) is an optimized approach that takes inputs from multiple weak models to yield a robust prediction. The proposed hybrid approach achieves a maximum accuracy of 95.2%, outperform other state-of-the-art techniques. Notably, this high accuracy is achieved using only three of the most significant parameters, showcasing the efficacy of the proposed methodology.

The main contribution of this work includes:

- Last 8 years' data (2013- 2021) is gathered from CPCB and converted into machine readable format for the further processing.
- WQI is calculated on 9 sites of Delhi on four parameters such as pH, DO, BOD, COD.
- Various models such as LR, NB, SVM, DT and XGBoost are applied.
- A hybrid approach based on LSA (Latent Semantic Analysis) and XGBoost is proposed based on various parameters of water.

Related work

This section presents the work done on prediction of WQI.

Ahmed et.al. [6] explored various techniques based on Four input parameters. The results depict that gradient boosting was most efficient in prediction of WQI, and the multi-layer perceptron attains highest WQC classification accuracy at 85.07%. This proposed methodology achieved significant accuracy using minimal parameters set.

In another work, Wang et.al. [9] Focused on model stacking approach. Microbial contamination in beach water poses risks to swimmers due to exposure to harmful pathogens. An ensemble approach known as model stacking was proposed for water quality assessment for beaches. Outputs from five machine learning models were fed as an input to another model. In this, accuracy rankings for the stacking model remained consistent for first two years, with average accuracy of 78%, 81%, and 82.3% respectively. Silberg et al. [10] utilized an approach that combined attribute-realization with SVM algorithm for Chao Phraya River. The study, based on a historical dataset spanning 2008 – 2019 and encompassing various parameters, followed a four-step process: data pre- processing, attribute evaluation, exploration of mathematical functions. The study observed that different combinations of attributes and mathematical functions resulted in

varied performance. Validation of the approach confirmed that proposed method proved to be a robust method for classifying river water quality, achieving an accuracy range of 0.86 to 0.95 when using three to six attributes. This underscores the effectiveness of the AR-SVM approach in accurately categorizing the Chao Phraya River's water quality based on diverse attributes.

Yilma et al. [11] sought to present a comprehensive assessment of pollution levels in The Little Akaki River. The approach employed neural network on twelve parameters gathered from 27 sites. The results indicated that, with the exception of one upstream site, all sampling locations were classified under the poor water quality category.

In their study, Bui et al. [12] employed both standalone algorithms and data-mining algorithms on Iran Water Quality Index using six years of monthly data (2012 to 2018) in the Talar catchment. Hybrid algorithms demonstrated an enhancement, although this improvement was not uniform across all cases.

In their work, Ding et al. [13] introduced a hybrid intelligent algorithm. The initial application of PCA serves to reduce data dimensionality by compressing 23 factors into 15 indices following by Genetic algorithm to enhance the dimensions of the BPNN. The results attain an overall prediction rate of approximately 91%.

In their study, Azad et al. [14] explored nature inspired and fuzzy systems to predict water quality in Gorganroud River water. ANFIS-DE model in accurately predicting Electrical

Conductivity and Total Hardness in Gorganroud River water.

Zhang et al. [15] introduced a hybrid model named HANN, to anticipate the overall performance of Drinking Water across China. The approach utilized monthly data from 45 DWTPs. The resulting HANN model demonstrated excellent performance in simulating training datasets, exhibiting enhanced predictive accuracy.

Further, Hassan et al. [8] employed several techniques to classify water quality across diverse locations in India. The previous studies are compared and are presented in Table 1.

Table 1. Comparative analysis of existing state-of-the-art techniques

Author and Year	Machine learning model used	Dataset Used	Water_ Parameters	Evaluation Metrics	Results
Ding et.al., 2014, [13]	PCA GA ,BPNN	River water	Total 23 Factors aggregated into 15 parameters.	Accuracy	Total Overall prediction rate =91%
Yilma et.al., 2018 [11]	ANN	Little Akaki River	12 water parameters	R2	R2 of 0.95 was attained
Azad et.al. 2018. [14]	GA, Ant Colony Optimization and Differential Evolution	Gorganroud River water	Electrical Conductivity, Sodium Absorption Ratio,Total Hardness	R2, RMSE, MAPE	ANFIS exhibited the best performance
Ahmed et.al., 2019, [6]	Polynomil regression, GB, MLP	-	Temperature, Turbidity, pH, TDS	MAE, Accuracy	MAE of 1.9642 and 2.7273 for WQI Accuracy= 85.07%
Zhang et.al., 2019 [15]	Hybrid Statistical Model HANN, Integrating, ANN and GA	DWTPs across China	Temperature, Chemical Oxygen EC,CE	Mean Squared Error	Results indicated a close connection between DWTP water production and wáter quality and operational parameters

Author and Year	Machine learning model used	Dataset Used	Water_ Parameters	Evaluation Metrics	Results
Bui et.al., 2020 [12]	RF, M5P, Data-Mining Algorithms	Talar catchment of Iran	All wáter quality parameters	Pearson correlation coefficients	FC and TS had the greatest and least impact.
Wang et.al. 2021 [9]	Model Stacking	Beaches Dataset	Dissolved Solid, pH, Temperature, BOD	Accuracy	Accuracy of 78%, 81%, and 82.3% respectively.
Solberg et.al., 2021, [10]	AR and SVM	Chao Phraya river	NH3-N, TCB, FCB, BOD, DO, and Salanity	Accuracy	Attained an accuracy of 0.86-0.95.
Hassan et.al., 2021, [8]	RF, NN, MLR, SVM, and BTM	Various locations in India	DO, BOD, EC	Kappa coefficient, Accuracy	The results highlighted several influencing factors.

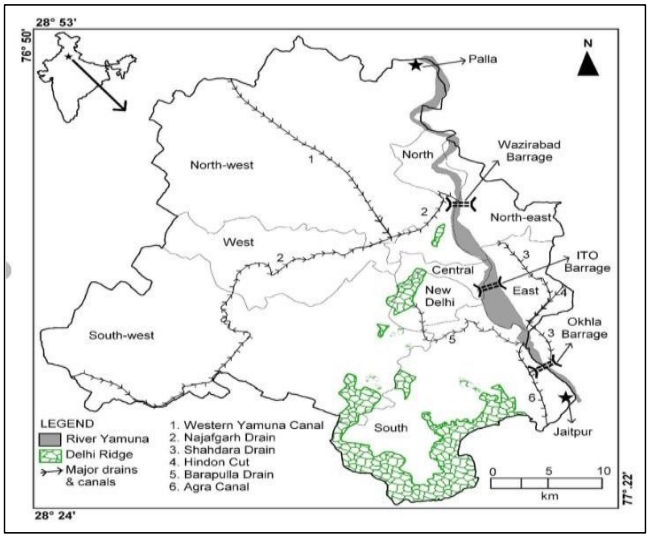
2. Methodology

The proposed methodology is divided into various steps. Steps are explained below in detail;

Dataset Collection

To perform the data analysis, data is gathered from the government Central Pollution Control Board for different locations of Delhi Region. Data is provided for 9 regions/area for 4 parameters for the years 2013 to 2021. Various location is represented trough L1 to L9. The 9 locations of Yamuna River are presented in fig 1.

Fig 1. Yamuna River Water Stations



Data Preprocessing

For the collected dataset, data is pre-processed by checking for missing values. Then, all the water quality parameters are normalized using min-max normalization approach. Then, normalized parameters are passed further for computation of Water Quality Index and further processing [16-17].

Water Quality Index (WQI)

Based on these four parameter presented in Table 2, WQI is calculated at each location from year 2013-2021 [18-19]. WQI is calculated using equations (1)-(4).

$$\text{Water Quality Index(WQI)} = \frac{\sum_{i=1}^{i=n} Wiqi}{\sum qi} \quad (1)$$

$$\text{Normalized value of each parameter} = \left(\frac{Measured\ Value - Min\ Value}{Max\ Value - Min\ Value}\right) * 100 \quad (2)$$

$$\text{Sub Index of each Parameter} = (Normalized\ Value * Weight) \quad (3)$$

$$\text{WQI} = \frac{\sum Sub\ Index\ of\ Each\ Parameter}{\sum weights} \quad (4)$$

The standard values the WQI is as per CPCB Delhi depicted in Table 2.

Table 2. WQI values and its Classification

WQI Range	WQI Classification
0-25	Excellent
26-50	Good
51-75	Poor
76-100	Very Poor
Above 100	Not fit for Drinking

Fig.2 presents the framework of the proposed approach and various components of the framework are that are explained further

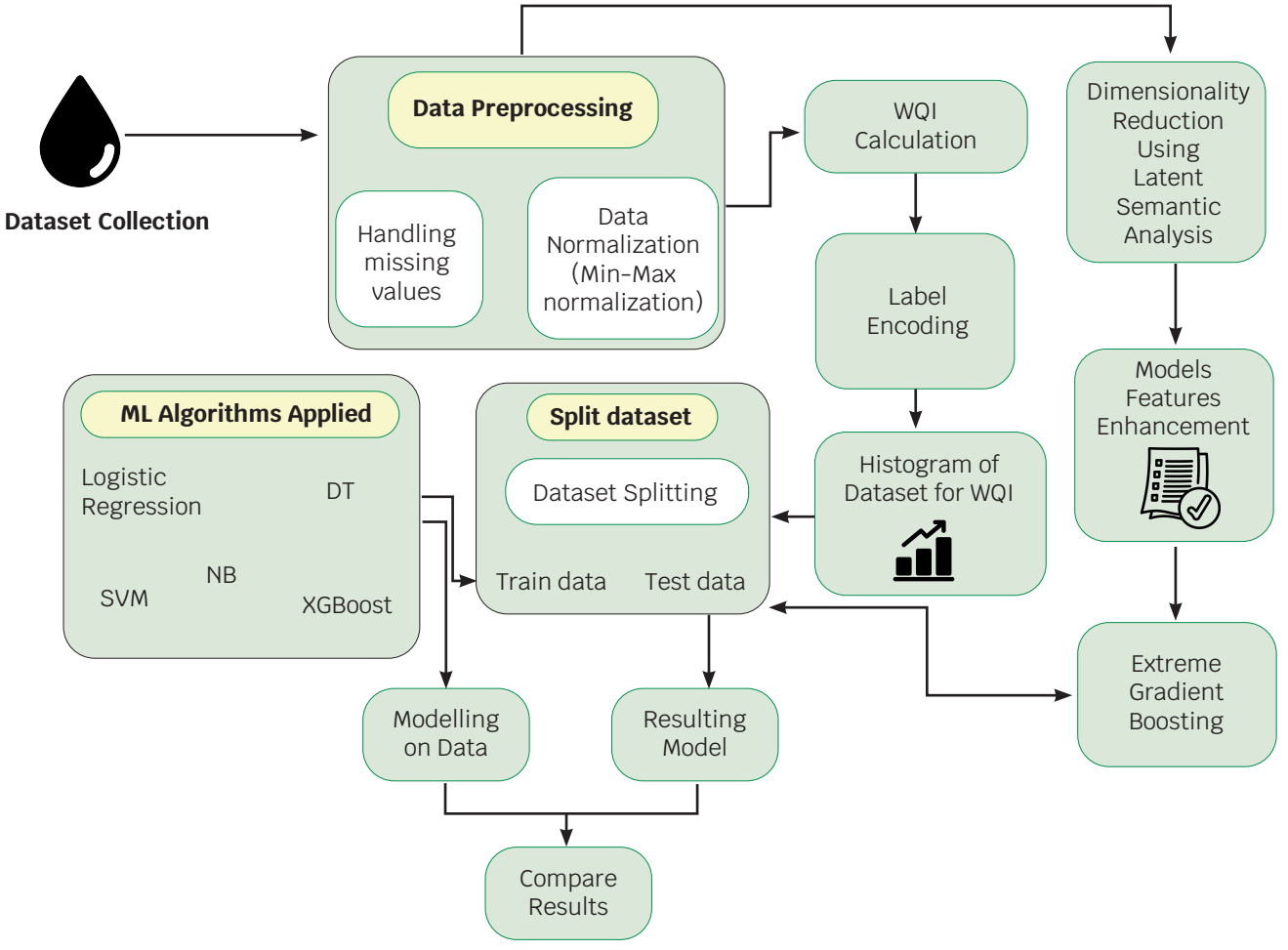


Fig. 2. Framework of the proposed methodology

Machine Learning Models

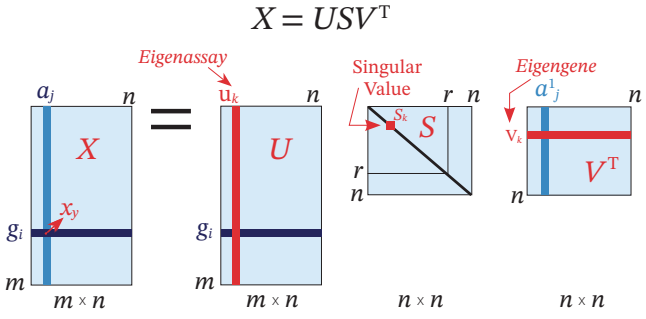
The dataset is now further split based on 70:30 ratios. On the train data, several models namely are applied. After Training the models on train data, the models are tested and WQI values are predicted. Further, the results are compared based on various evaluation metrics.

Proposed Methodology

After prediction of Water Quality Index, a new hybrid approach is proposed for improved results. In this hybrid methodology, after pre-processing of data, Laten Semantic Analysis is applied. Latent Semantic Analysis is used for Dimensionality reduction, which will further enhance the features or parameters of water i.e., BOD, COD, pH and Temperature. Latent Semantic Analysis works on the principle of Singular Vector Decomposition (SVD). It is applied to numerical parameters that involves leveraging the technique's ability to identify and enhance latent patterns, leading to a more informative feature representation. This enhanced representation can contribute to better insights and improved performance in prediction tasks. To ensure consis-

tency in scale, the numerical data undergoes normalization before the application of Latent Semantic Analysis (LSA). Normalization is a critical step to enhance the effectiveness of LSA. Subsequently, Singular Value Decomposition is implemented on the term-document matrix. Following SVD, only the top k singular values and their corresponding columns in U and V matrices are retained. This selective retention reduces the dimensionality of the data while preserving the most. The working of LSA is represnted in Fig.3.

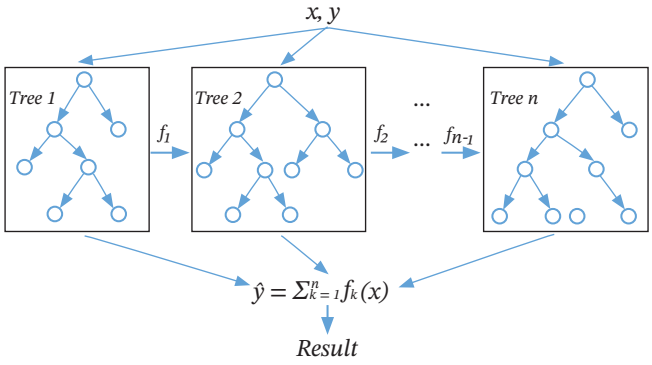
Fig.3. Process of Latent Semantic Anlaysis



In this, the reduced U matrix serves as a transformed representation of the original features, capturing latent semantic relationships between them, thus, an enhanced set of features is created that encapsulates the underlying structure and relationships within the numerical data. These features may highlight latent patterns or relationships in the numerical data that were not apparent in the original feature set. After feature enhancement, the data is reconstructed by multiplying the reduced U, Σ, and V^T matrices. The reconstructed matrix represents an approximation of the original data with the enhanced features [20-22].

Therefore, the improved features obtained through Latent Semantic Analysis (LSA) can be utilized for the subsequent training of Extreme Gradient Boosting which is an ensemble learning technique and constructs a robust predictive model. The mathematical model behind the XGBoost involves the iterative addition of weak learners to the ensemble while optimizing an objective function as shown in fig 4.

Fig.4. Working of Extreme Gradient Boosting



The objective function is the sum of the loss function over all training instances and a regularization term as depicted in Fig. 4. The dataset with enhanced features (X_enhanced) is splitted. An XGBoost model is initialized with parameters like the objective (classification), number of boosting rounds (n_estimators), maximum tree depth (max_depth) and learning rate [23-25].

Pseudocode of the proposed approach

Algorithm1: For WQI and Class Assignment

Input: A numerical dataset represented as a matrix X havingn parameters ***Output*: *WQI*, *Target Variable y* (*class*)**
Step 1: Read the matrix f *matrix = read_csv("sample matrix.csv")*
Step 2: Handle missing values *X_processed = handle_missing_values(matrix)*

Input: A numerical dataset represented as a matrix X havingn parameters ***Output*: *WQI*, *Target Variable y* (*class*)**
Step 3: Calculate Sub Index *sub_index_values calculate_sub_index(X_processed*
Step 4: *Calculate Water Quality Index (WQI) based on sub-indexwqi_values = calculate_wqi(sub_index_values)*
Step 5: *Assign class labels based on standard WQI val* class_labels = assign_class_labels(wqi_values) ***Function***
Function handle_missing_values(X): # Handle missing values (emedian) X_processed = impute_missing_values(X) return X_processed ***function calculate_sub_index(X): sub_index_values*** = (Normalized Value*Weight) return sub_index_values *function calculate_wqi(sub_index_values):* $\sum Sub\ Index\ of\ Each\ Parameter$ *return Wqi_values*
function assign_class_labels(wqi_values) # Assign class labels values *class_labels = classify_wqi(wqi_values)* *return class_labels*

Algorithm 2: Hybrid Approach and Class Assignment for unknown parameters

Input: Assume a numerical dataset represented as a matrix X having parameters and a Class label y ***Output:Target Varible y (class)for unknown parameters***
Step 1: Read the matrix *matrix = read_csv("matrix1.csv")*
Step 2: *Pre-processing and Normalize the data X_normalized = normalize(matrix)*
Step 3: Apply Singular Value Decomposition (SVD) *U, Sigma, Vt = svd(X_normalized)*
2.1. Step 4: Choose the number of components (k) to retain *k = choose_k()*
2.1. Step 5: Retain the top k components *U_k = U[:, :k]* *Sigma_k= Sigma[:k] Vt_k = Vt[:k, :]*
2.1. Step 6: Feature enhancement *X_enhanced = U_k * Sigma_k * Vt_k*
Step 7: Split the data into training and testing sets *X_train, X_test, y_train, y_test = train_test_split(X_enhanced, y, =42)*
Step 8: Initialize and configure the XGBoost model **xgb_model = XGBClassifier (objective='binary:logistic', num_rounds=100, max_tree_depth=3, learn_rate=0.1, sampling_rate=0.7, tree_colsample=0.7, seed_value=42)**
Step 9: Train the XGBoost model on the training data **xgb_model.fit(X_train, y_train)**

Input: Assume a numerical dataset represented as a matrix X having parameters and a Class label y ***Output:Target Varible y (class)for unknown parameters***
Step 10: Make predictions on the test set y_pred = xgb_model.predict(X_test)
Step 11: Evaluate the performance of the model accuracy = accuracy_score(y_test, y_pred) precision = precision_score (y_test, y_pred) recall = recall_score(y_test, y_pred)
Step 12: Predict for new parameters **Class_label= label_class(parameters)**

3. Implementation

The primary intent is to evaluate quality of water by considering multiple parameters in the proximity of industrial areas and along the course of the Yamuna River in Delhi. samples from various points along the river were collected, with a specific emphasis on locations near industrial establishments. The gathered water samples undergo comprehensive analysis for key water quality parameters. Following the data collection and analysis phase, the WQI is computed [26-27]. Subsequently, various machine learning models are employed to further refine the WQI calculations. Additionally, a novel hybrid approach incorporating Latent Semantic Analysis (LSA) and the XGBoost machine learning model is proposed to enhance predictive accuracy.

To ensure a robust dataset, information is sourced from the Central Pollution Control Board (CPCB), a governmental body, encompassing different locations within the Delhi región [26]. The data spans nine distinct regions or areas, denoted as L1 to L9, and covers four essential parameters for the years 2013 to 2021[27]. This meticulous approach allows for a comprehensive understanding of variations across the specified regions and parameters, facilitating a nuanced analysis of the environmental dynamics in this critical area. The Sample data given by the authority in image form which was converted into excel/csv format for further processing as presented in Table 3.

Table 3. Various parameter of water pollution at 9 locations of Yamuna River Delhi in March 2023

Location	Location_ Represented As	pH	COD(mg/l)	BOD (mg/l)	DO (mg/l)
Palla	L1	8.3	8	2	9.0
Surghat	L2	8	12	2.5	3.8
Khajori Paltoon	L3	7.9	112	28	NIL
Kudesia Ghat	L4	7.8	80	24	NIL
ITO Bridge	L5	8.1	72	24	1.3
Nizamaadin Bridge	L6	8.0	72	23	1.2
Agar Canal (Okhla)	L7	7.9	96	32	NIL
Shahdara Drains (Downstream Okhla Drain)	L8	7.8	112	36	NIL
Agra Canal	L9	8	96	30	NIL

WQI is computed by considering these parameters. The reference values for Wi and Qi for the Yamuna River are obtained from the Central Pollution Control system, as outlined in Table 4. The minimum and maximum values for the nine locations, according to the CPCB, are presented in Table 5. Utilizing these parameters, the WQI is computed and presented below.

Table 4. Standard values of various water quality parameters

S.No	Parameter	Standard Value	Weighted Value
1.	pH	6.5-8.5	0.2272
2.	COD	0-3	0.0077
3.	DO	5	0.3862
4.	BOD	3	0.3213

Table 5. Lowest and Highest values of 9 locations of Yamuna River

Location	pH		COD		BOD		DO	
	Min	Max	Min	Max	Min	Max	Min	Max
L1	6.7	8.6	1	6	1	14	3.9	9.6
L2	7.3	8.5	1.2	6.7	2.5	11	4.6	14
L3	6.6	7.4	1.6	7.2	8	10	1.9	8.3
L4	7.2	7.5	1.4	6.8	34	38	0.3	1.6
L5	7.4	7.6	2	6.5	26	62	0.3	1.8
L6	7.3	8	2.1	6.3	22	48	0.3	3.2
L7	7.4	7.9	2.3	6.9	22	56	0.3	2.9
L8	7.4	8	1.9	7	38	83	0.3	2.2
L9	7.3	8.1	2.0	7.1	37	76	0.28	2.1

Fig 5. Box plots for all the water quality parameters

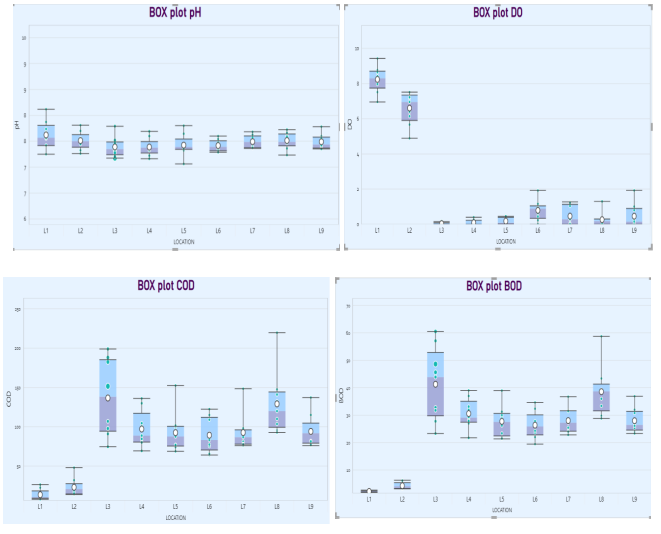


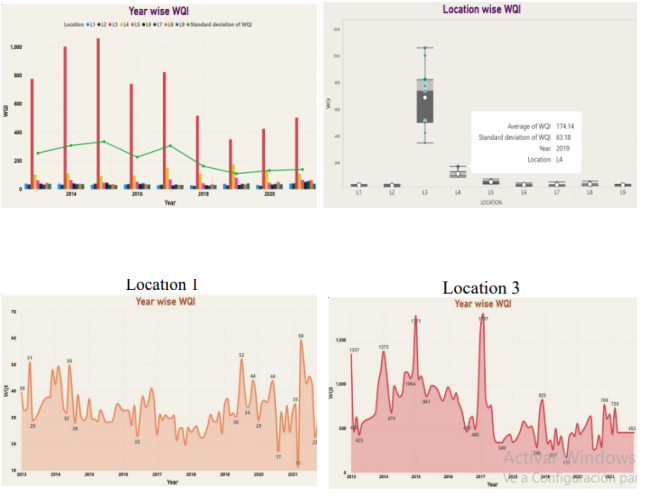
Figure 5 illustrates the Box plot representing various parameters. Upon thorough analysis of the Box plot, it is evident that L1 and L2 exhibit the most favorable water quality parameters. While Location 1 (Palla) displays slightly elevated COD levels, all other parameters conform to standard observations. Conversely, at Location 2, COD levels are higher compared to L1. In contrast, Location 3 exhibits the least favorable water quality parameters among all locations, featuring a pH value below the standard threshold of 7.5, no recorded Dissolved Oxygen, an average BOD of 55.3, an average COD of 134, and a maximum COD value of 198.

In comparison, Location 8 and Location 9 demonstrate relatively superior water quality parameters when contrasted with Location 4, Location 5, and Location 6. The authors calculated the Water Quality Index for each location from 2013 to 2021 based on these four parameters, using equations 1, 2, 3, and 4. The average values of WQI for each year are calculated, and Table 6 presents the average WQI based on the year for each location. To facilitate better comprehension and analysis, a bar graph and box plot are provided in Fig.6.

Table 6. Average WQI for 8 years for all locations

Year	LOCATIONS								
	L1	L2	L3	L4	L5	L6	L7	L8	L9
2013	38.53	32.91	776.42	103.17	62.49	38.68	32.40	44.67	36.94
2014	37.88	30.83	1004.09	110.44	63.64	40.75	34.66	38.24	34.95
2015	31.09	38.31	1062.12	93.18	45.56	44.72	31.17	35.48	27.91
2016	33.00	35.31	741.36	95.42	52.35	36.48	40.96	35.85	32.03
2017	27.67	31.47	824.33	151.54	67.85	26.20	33.14	29.87	29.39
2018	26.56	22.68	517.39	106.71	41.86	27.42	24.96	35.75	29.09
2019	37.62	26.91	350.23	174.14	80.43	29.59	37.45	33.43	40.41
2020	29.81	23.13	425.34	135.34	44.74	31.78	36.99	51.48	35.32
2021	38.78	40.58	504.05	107.32	65.08	50.55	57.70	63.82	37.77

Fig. 6 Bar graphs and line charts representing WQI



The examination underscores that the water quality in the National Capital Region (NCR) falls short of meeting acceptable standards. While there has been

a marginal enhancement in water quality post the COVID-19 pandemic, it still does not align with the prescribed standard value of Water Quality Index (WQI). Based on classifications given by table 2, further classes are grouped into three classes to enhance the accuracy as depicted in Table 7.

Table 7. Classification of WQI Range

WQI Range	WQI Classification
0-50	1
51-100	2
Above 100	0

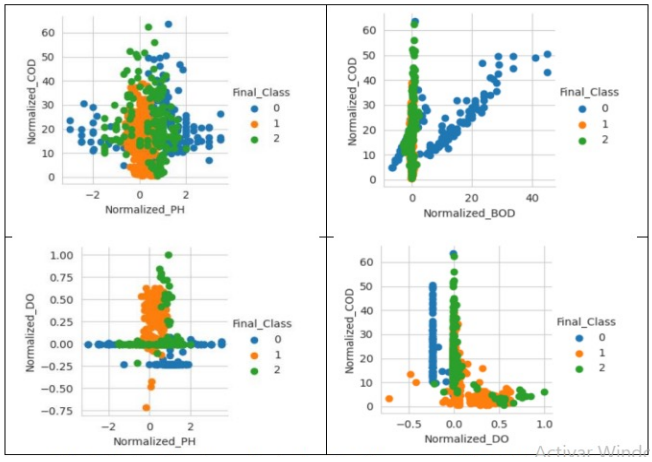
For further processing the sample data is prepared having 4 parameters and class label as shown in Table 8.

Table 8. Sample data for processing into machine learning models

Month	Location	pH	COD	BOD	DO	Quality	Class
1/1/2013	L1	7.2	32	3	10.2	Good	1
2/1/2013	L1	7.4	20	2	8.8	Good	1
3/1/2013	L1	7.4	24	2.4	8.5	Good	1
4/1/2013	L1	7.7	32	3	11.5	Poor	2
5/1/2013	L1	7.4	16	2.2	7.7	Good	1
10/1/2013	L1	8	20	2.1	7.7	Good	1
11/1/2013	L1	7.5	16	2.6	9.5	Good	1

Subsequently, the data is normalized, and a scatter plot is presented in Fig. 7 to illustrate the correlation between the normalized values and the class labels across various classes.

Fig. 7. Scatter plot of normalized values for all parameters BOD, COD, DO, Ph



These normalized values presented in Fig. 8 are subsequently condensed to three components through the application of the SVD and LSA model. The values for these components are presented in Fig.9. Three components are defined as [0,1,2] and first five values are shown.

Fig.8. Normalized values for all four water quality parameters BOD, COD, DO, pH

	0	1	2	3	4	5	6	7
normalized_PH	0.263158	0.368421	0.368421	0.526316	0.368421	0.684211	0.421053	0.578947
normalized_COD	6.200000	3.800000	4.600000	6.200000	3.000000	3.800000	3.000000	3.800000
normalized_BOD	0.153846	0.076923	0.107692	0.153846	0.092308	0.084615	0.123077	0.092308
normalized_DO	0.623762	0.485149	0.455446	0.752475	0.376238	0.376238	0.554455	0.752475

4 rows x 840 columns

Fig. 9. Enhance features obtained after applying LSA

	0	1	2
0	-0.891675	-0.231559	0.388967
1	-0.891886	-0.287322	0.349263
2	-0.894477	-0.285546	0.344056
3	-0.868946	-0.334442	0.364803
4	-0.897484	-0.294297	0.328498

Then these normalized and reduced values are passed to XG BOOST machine learning models for further predication

4. Results

To implement the machine learning model, as elaborated earlier, the samples are partitioned into three distinct classes. Class 1 signifies good water quality, Class 2 denotes poor quality, and Class 0 indicates water unfit for drinking. In total, we have 840 samples containing Date, location, pH, DO, BOD, COD, and Class label information. Colab, a Google Python environment, is employed for file reading and applying machine learning models. The authors utilized various libraries, including NumPy, Pandas, Scikit-learn, and Seaborn for plotting. The dataset consists of 465 records for Class 1, 199 records for Class 2, and 176 records for Class 0, displaying a roughly balanced distribution. The authors trained the model based on this data.In the proposed hybrid approach, all data parameters are initially normalized, and Latent Semantic Analysis (LSA) is employed for dimension reduction. While the authors used four parameters in this instance, dimension reduction can be extended to include more parameters. Following normalization and LSA, the parameters undergo training for machine learning models. Various algorithms, such as Logistic Regression [28-29], Decision Tree [30], Support Vector Machine, Naïve Bayes, and XGBoost, are applied, alongside the proposed hybrid method. The data is splitted and 10-fold cross- validation is implemented across all algorithms. The applied models are validated through Precision, Recall, and Accuracy metrics. Precision, Recall and accuracy is computed using these equations [31-33].

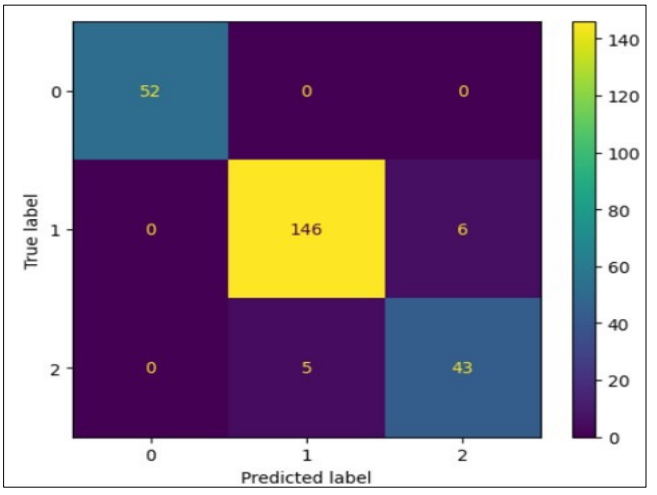
Confusion Matrix for the proposed approach is attained in Fig. 10.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (5)$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (6)$$

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predication}{Total\ Number\ of\ Predication} \quad (7)$$

Fig.10. Confusion matrix attained for the proposed approach



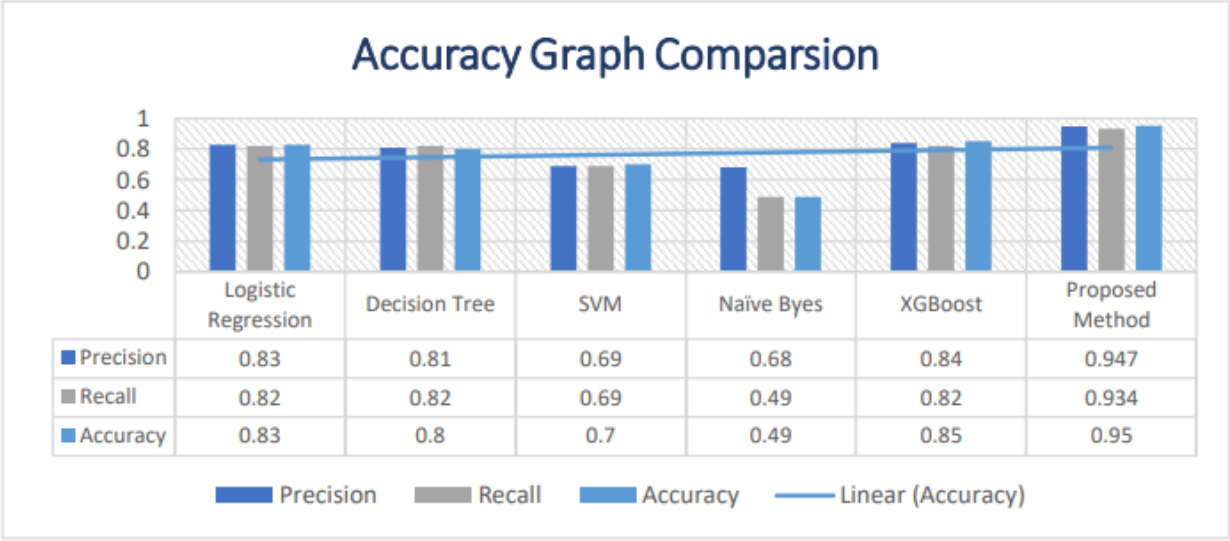
The conclusive outcomes are presented in Table 10. Notably, the proposed approach achieves the highest accuracy at 0.95 [34]. The results are depicted in Fig.11.

Table 10. Comparison of various techniques based on evaluation metrics

Algorithm	Precision	Recall	Accuracy/NOR Accuracy
Logistic Regression	0.83	0.82	0.83
Decision Tree	0.81	0.82	080
SVM	0.69	0.69	0.70
Naïve Byes	0.68	0.49	0.49
XGBoost	0.84	0.82	0.85
Proposed Method	0.947	0.934	0.95

Precision, Recall, and Accuracy metrics. Precision, Recall and accuracy is computed using these equations [31-33].

Fig.11. The comparison of machine learning models and proposed approach



5. Discussion

In our constant search for innovative and efficient approaches in the field of machine learning, we have identified a recurring pattern in various researches addressing water quality index (WQI) prediction. Researches such as Ahmad et al. [6], Sakizadeh [9], Gazzaz et al. [11], Parmar and Bhardwaj [35] and Adnan et al. [36] have adopted machine learning methods that use large sets of parameters to make these predictions. While this approach can be effective in obtaining results, it presents significant challenges in practice. In particular, the complexity and cost of implementing systems that handle multiple parameters in real time can be prohibitive, limiting their applicability in environments where economic efficiency is crucial [35].

Given this situation, we have developed a unique methodology that significantly simplifies the prediction process. Through data normalization and the implementation of latent semantic analysis (LSA) for dimension reduction, we have managed to make accurate predictions using only four water quality parameters. This distinctive approach has allowed us to achieve a peak accuracy of 95%, a result that not only rivals, but surpasses, the performance of the most advanced methods available in the literature. For example, historical results show that the best accuracy obtained by a multilayer perceptron, which used ten parameters, was 91%. This comparison highlights not only the superiority of our methodology in terms of accuracy, but also its ability to optimize water quality prediction models.

Furthermore, our strategy not only focuses on improving accuracy, but also addresses the need for practical solutions for real-time water quality monitoring systems. The reduction in the number of parameters required to make accurate predictions opens the door to the implementation of more affordable and sustainable systems. Consequently, our approach not only contributes to the advancement of knowledge in the field of machine learning applied to water quality, but also facilitates the creation of monitoring tools that are accessible and efficient, thus driving positive change in the way water resources are managed and monitored. This development therefore marks a significant step towards the integration of more effective monitoring technologies in the environmental field.

6. Conclusion:

This research delves into the intricate nexus between water quality, environmental health and ecosystem

vitality of the Yamuna River, an essential water resource facing serious threats. The alarming levels of pollution, predominantly driven by industrial discharges and urban waste, underscore the urgent need for robust and effective measures to safeguard this vital resource. Water quality not only impacts aquatic biodiversity but also has direct implications on the health of communities that depend on it for their livelihoods.

The traditional approach of calculating the Water Quality Index (WQI), while critical to understanding the current situation, presents significant challenges such as time-consuming data collection processes and the associated rising financial costs. These limitations can hamper rapid and effective response to the water quality crisis. Recognizing these constraints, the study embarks on a pioneering journey into the realm of machine learning, a field that presents unprecedented opportunities to improve our predictive capabilities in this context.

The proposed hybrid approach, which integrates Latent Semantic Analysis (LSA) and Extreme Gradient Boosting, emerges as a beacon of innovation. By reducing the dimensionality of the data and improving the representation of relevant features, this methodology not only streamlines the WQI prediction process but also achieves an impressive accuracy of 95.2%. This result is a testament to the potential of advanced predictive models to address the evolving complexities of water quality dynamics.

The major contributions of this work include eight years of exhaustive data collection, water quality index calculations at critical sites, and the introduction of a novel hybrid approach that improves our understanding of the water status in the Yamuna River. Furthermore, this approach sets a precedent for future research in this domain by providing a model that can be adapted and applied to other threatened water bodies. As we face the challenges of an ever-changing environment, this research serves as a guiding light, illuminating the path towards sustainable water management and the preservation of vital aquatic ecosystems such as the Yamuna River, thereby promoting a healthier and more balanced future for all.

7. References

[1] Mohtar, W. H. M. W., Maulud, K. N. A., Muhammad, N. S., Sharil, S. & Yaseen, Z. M. (2019). "Spatial and temporal risk quotient based river assessment for water resources management". *Environmental Pollution*, 248, 133-144.

[2] Meybeck, M. (1976). "Total mineral dissolved transport by world major rivers/Transport en sels dissous des plus grands fleuves mondiaux". *Hydrological Sciences Journal*, 21(2), 265-284. https://doi.org/10.1080/02626667609491631

[3] Sunil, C., Somashekar, R. K. & Nagaraja, B. C. (2010). "Riparian vegetation assessment of Cauvery River basin of South India". *Environmental Monitoring and Assessment*, 170, 545-553. https://doi.org/10.1007/s10661-009-1256-3

[4] Rani, M., Akolkar, P. & Bhamrah, H. S. (2013). "Water quality assessment of River Yamuna from origin to confluence to River Ganga, with respect to biological water quality and primary water quality criteria". *Journal of Entomology and Zoology Studies*, 1(6), 1-6.

[5] Misra, A. K. (2010). "A river about to die: Yamuna". *Journal of water resource and protection*, 2(5), 489.

[6] Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., ... & Elshafie, A. (2019). "Machine learning methods for better water quality prediction". *Journal of Hydrology*, 578, 124084.

[7] Chia, S. L., Chia, M. Y., Koo, C. H. & Huang, Y. F. (2022). "Integration of advanced optimization algorithms into least-square support vector machine (LSSVM) for water quality index prediction". *Water Supply*, 22(2), 1951-1963.

[8] Hassan, M. M., Hassan, M. M., Akter, L., Rahman, M. M., Zaman, S., Hasib, K. M., ... & Mollick, S. (2021). "Efficient prediction of water quality index (WQI) using machine learning algorithms". *Human-Centric Intelligent Systems*, 1(3-4), 86-97.

[9] Wang, L., Zhu, Z., Sassoubre, L., Yu, G., Liao, C., Hu, Q. & Wang, Y. (2021). "Improving the robustness of beach water quality modeling using an ensemble machine learning approach". *Science of The Total Environment*, 765, 142760.

[10] Sillberg, C. V., Kullavanijaya, P. & Chavalparit, O. (2021). "Water quality classification by integration of attribute-realization and support vector machine for the Chao Phraya River". *Journal of Ecological Engineering*, 22(9), 70-86.

[11] Yilma, M., Kiflie, Z., Windsperger, A. & Gessese, N. (2018). "Application of artificial neural network in water quality index prediction: a case study in Little Akaki River, Addis Ababa, Ethiopia". *Modeling Earth Systems and Environment*, 4, 175-187.

[12] Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H. & Kazakis, N. (2020). "Improving prediction of water quality indices using novel hybrid machine-learning algorithms". *Science of the Total Environment*, 721, 137612.

[13] Ding, Y. R., Cai, Y. J., Sun, P. D. & Chen, B. (2014). "The use of combined neural networks and genetic algorithms for prediction of river water quality". *Journal of applied research and technology*, 12(3), 493-499.

[14] Azad, A., Karami, H., Farzin, S., Saeedian, A., Kashi, H. & Sayyahi, F. (2018). "Prediction of water quality parameters using ANFIS optimized by intelligence algorithms (case study: Gorganrood River)". *KSCE Journal of Civil Engineering*, 22, 2206-2213.

[15] Zhang, Y., Gao, X., Smith, K., Inial, G., Liu, S., Conil, L. B. & Pan, B. (2019). "Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network". *Water research*, 164, 114888.

[16] Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A. A., Mohamed, A. & Ashraf, I (2022). " Water quality prediction using KNN imputer and multilayer perceptron". *Water*, 14(17), 2592.

[17] 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.

[18] Akhtar, N., Ishak, M. I. S., Ahmad, M. I., Umar, K., Md Yusuff, M. S., Anees, M. T., ... & Ali Almanasir, Y. K. (2021). "Modification of the water quality index (WQI) process for simple calculation using the multi-criteria decision-making (MCDM) method: a review". *Water*, 13(7), 905.

[19] Călmuc, V. A., Călmuc, M., Țopa, M. C., Timofti, M., Iticescu, C., & Georgescu, L. P. (2018). "Various methods for calculating the water quality index". *Analele Universităţii" Dunărea de Jos" din Galaţi. Fascicula II, Matematică, fizică, mecanică teoretică/Annals of the" Dunarea de Jos" University of Galati. Fascicle II, Mathematics, Physics, Theoretical Mechanics*, 41(2), 171-178.

[20] Pau, F. & Pablo, G. B. (2023). *Revisiting the Probabilistic Latent Semantic Analysis: The Method, Its Extensions and Its Algorithms*.

[21] Hutchison, P. D., George, B. & Guragai, B. (2023). "Application of Latent Semantic Analysis in Accounting Research". *Journal of Information Systems*, 37(3), 139-155.

[22] Landauer, T. (2023, January). "Latent semantic analysis: theory, method and application". In *Computer Support for Collaborative Learning* (pp. 742-743). Routledge.

[23] Yan, T., Zhou, A. & Shen, S. L. (2023). "Prediction of long-term water quality using machine learning enhanced by Bayesian optimisation". *Environmental Pollution*, 318, 120870.

[24] Dritsas, E. & Trigka, M. (2023). "Efficient Data-Driven Machine Learning Models for Water Quality Prediction". *Computation*, 11(2), 16.

[25] Koh, J. (2023). " Gradient boosting with extreme-value theory for wildfire prediction". *Extremes*, 1-27.

[26] https://www.dpcc.delhigovt.nic.in/#gsc.tab=0

[27] https://cpcb.nic.in/yamuna-monitoring-committee/

[28] Goldar, B. & Banerjee, N. (2004). "Impact of informal regulation of pollution on water quality in rivers in India". *Journal of Environmental Management*, 73(2), 117-130.

[29] Kaushik, C. P., Sharma, H. R., Jain, S., Dawra, J. & Kaushik, A. (2008). "Pesticide residues in river Yamuna and its canals in Haryana and Delhi, India". *Environmental Monitoring and Assessment*, 144(1-3), 329-340.

[30] Mamais, D., Jenkins, D. & Pitt, P. (1993). "A rapid physical-chemical method for the determination of readily biodegradable soluble COD in municipal wastewater". *Water Research*, 27(1), 195-197.

[31] Mishra, A. K. (2010). "A river about to Die: Yamuna". *Journal of Water Resource and Protection*, 2(5) 489-500.

[32] Qureshimatva, U. M., Maurya, R. R., Gamit, S. B., Patel, R. D. & Solanki, H. A. (2015). "Determination of PhysicoChemical Parameters and Water Quality Index (Wqi) of Chandlodia Lake, Ahmedabad, Gujarat, India". *Journal of Environmental and Analytical Toxicology*, 3(3), 1176-1193.

[33] Suthar, S., Sharma, J., Chabukdhara, M. & Nema, A. K. (2010). "Water quality assessment of river Hindon at Ghaziabad, India: impact of industrial and urban wastewater". *Environmental Monitoring and Assessment*, 165(1-4), 103-112.

[34] Shah, M. I., Alaloul, W. S., Alqahtani, A., Aldrees, A., Musarat, M. A. & Javed, M. F. (2021). "Predictive modeling approach for surface water quality: development and comparison of machine learning models". *Sustainability*, 13(14), 7515.

[35] Parmar, K. S. & Bhardwaj, R. (2013). "Water quality index and fractal dimension analysis of water parameters". *International journal of environmental science and technology*, 10, 151-164.

[36] Adnan, R. M., Mostafa, R. R., Elbeltagi, A., Yaseen, Z. M., Shahid, S. & Kisi, O. (2022). "Development of new machine learning model for streamflow prediction: Case studies in Pakistan". *Stochastic Environmental Research and Risk Assessment*, 1-35.

Consent to publish

The authors have read and approved the final manuscript.

Conflict of interest

The authors declare that they have no conflict of interest. This document reflects their views only and not those of the institution to which they belong.

Neetu Guptaa:

Research Fellow, Faculty of Engineering and Technology, Career Point University, Kota, India

Surendra Yadavb:

Department of Computer Science and Applications, Vivekananda Global University, Jaipur, India

Neha Chaudharyc:

Department of Computer Science and Engineering, Manipal University, Jaipur, India