Techno-computing Journal 1(3) 1–15(2025)
https://doi.org/10.71170/tecoj.2025.1.3.1-15

Research Paper

Advanced Data Driven Prediction of BOD in the Ganga River Using Multivariate Regression and Nonlinear Bilayered Neural Network Ensembles

Usman U. Aliyu[1,2]*, Abdulhayat M. Jibrin[3], Abubakar S. Baba[1] Ismail A. Mahmoud[4], Sukalpaa Chaki[1], Rakesh Kumar[1]

[1]Department of Civil Engineering, Federal University Dutsin-Ma, Katsina State
[2]Department of Civil Engineering, Sharda University, Greater Noida, India
[3]Department of Civil Engineering, King Fahad University of Petroleum and Minerals
[4]Department of Physics, Northwest University, Kano State
Corresponding Author Email: ualiyu12@fudutsinma.edu.ng

Abstract

Accurate prediction of Biochemical Oxygen Demand (BOD) is essential for understanding pollution dynamics and supporting effective water quality management in the Ganga River. This study develops a comprehensive data-driven modeling framework that integrates multivariate regression models with neural network ensemble techniques to forecast BOD concentrations using physiochemical and microbial water quality indicators. Four regression models, including Fine-Tree Linear Regression (FLR), Interactive Linear Regression (ILR), Robust Linear Regression (RLR), and Stepwise Linear Regression (SWLR), were developed using combinations of dissolved oxygen (DO), pH, conductivity, total coliform (TC), and fecal coliform (FC). Correlation analysis revealed moderate positive associations of BOD with pH (r = 0.26), conductivity (r = 0.23), and dissolved oxygen (r = 0.06), on the other hand, the microbial indicators showed weak negative correlations, indicating the need for advanced modeling frameworks beyond simple linear relationships. Model evaluation based on MSE, RMSE, MAE, and SMAPE showed that FLR models outperformed other regression models, with FLR-4 producing the lowest testing errors (MSE = 0.0043; RMSE = 0.0657) among all linear regressors. However, integrating the regression outputs into neural network ensembles significantly enhanced prediction accuracy. The Bilayered Neural Ensemble (BNE) models consistently performed best, with BNE-RLR (testing MSE = 0.0015; RMSE = 0.0381) and BNE-ILR (testing MSE = 0.0015; RMSE = 0.0392) providing the highest accuracy and stability across all performance indices. The findings demonstrate that coupling multivariate regression with neural network ensemble modeling provides a robust and highly accurate framework for BOD prediction in the Ganga River and other similar river systems.

Keywords: Biochemical Oxygen Demand; Ganga River; Machine Learning; Neural Network Ensembles; Water Quality

1. Introduction

The Ganga River is one of the world's most densely utilized freshwater systems and continues to face critical pollution pressures due to rapid urban expansion, untreated municipal sewage, industrial effluents, and intensified agricultural activities. These stressors have led to significant deterioration in water quality, particularly in stretches around major urban centers such as Kanpur, Varanasi, and Patna [1][2]. Among the various parameters used to assess aquatic health, BOD is widely recognized as a key indicator of organic pollution, ecosystem stability, and microbial oxygen consumption within the river environment [3]. Effective prediction of BOD is fundamental to pollution forecasting, regulatory

(Received 11Nov 2025; revised 01 Dec 2025; accepted 21 Dec 2025; first published online 24 Dec 2025); Journal homepage https://technocomputing.org/index.php/tecoj

planning, and sustainable river basin management across the Ganga basin. Despite its importance, modelling BOD remains methodologically challenging due to the nonlinear, heterogeneous, and dynamic interactions among physicochemical and microbial variables. Past studies indicate that relationships involving DO, pH, electrical conductivity, TC, and FC often deviate from purely linear patterns, especially under variable hydrological regimes [4] [5]. Traditional linear approaches have captured broad trends but fail to represent higher-order interactions and uncertainty, particularly during periods of high organic loading or monsoon-driven fluctuations [6]. Consequently, recent research has shifted toward data-driven models, including neural networks, fuzzy systems, and ensemble learning, which provide better representation of temporal variability and contaminant dynamics [7] [8].

Existing machine-learning studies on river systems have primarily relied on standalone models such as ANN, SVR, random forests, or fuzzy inference systems. While these methods exhibit improved predictive capability, they often lack interpretability or robustness when applied across large, spatially diverse datasets [9] [10]. Similarly, [11] study implements explainable machine learning for assessing groundwater quality, [12] used computational-based approaches and [13] applied supervised ML with ANN, but limited studies have explored hybrid frameworks that combine regression interpretability with ensemble predictive strength. Moreover, no prior work has systematically evaluated the integration of FLR, ILR, RLR, and SWLR within a unified neural network ensemble architecture for BOD forecasting in the Ganga River. This unexplored methodological gap motivates the present study. To address these limitations, this research develops a comprehensive hybrid modelling framework that couples four multivariate regression structures (FLR, ILR, RLR, and SWLR) with advanced neural network ensemble techniques. Regression models provide structured interpretability by quantifying relationships among BOD and selected water quality indicators, while neural ensembles, particularly BNE, enhance accuracy by reducing variance and stabilizing predictions across multiple resampled models. This dual-architecture approach aims to exploit the strengths of both modelling categories, thereby improving generalization performance and minimizing prediction errors that typically arise in standalone models. The application significance of this study extends to operational water quality management, real-time pollution surveillance, and early warning systems under the National Mission for Clean Ganga (NMCG). A hybrid regression ensemble framework offers a scalable tool for forecasting BOD with high precision, supporting decision-making for wastewater discharge control, treatment infrastructure planning, and environmental regulation enforcement. Beyond the Ganga basin, the proposed framework can be adapted for other polluted river systems facing complex contaminant dynamics. The outcomes may contribute to the broader scientific agenda of integrating interpretable statistical models with advanced machine-learning strategies to enhance predictive water quality analytics.

2. Methodology

2.1 Dataset and Study Area

The Ganga River basin spans a broad geographic range, approximately 21°06′ to 31°21′ N latitude and 73°02′ to 89°05′ E longitude. This study focuses on the river that spans Uttarakhand, Uttar Pradesh, Bihar, Jharkhand, and West Bengal, capturing its longitudinal variability through the influence of urban effluents, agricultural runoff, industrial discharges, and natural purification processes. Water-quality data were collected at key entry and exit points of each state, including notable locations such as upstream Jail Ghat and downstream Cremation Ghat (Bihar), Raj Mahal (Bihar), LCT Ghat (Jharkhand), Khagra–Beharampore and Diamond Harbour (West Bengal), Bijnor and Tarighat Ghazipur (Uttar Pradesh), and Sultanpur (Uttarakhand), as well as NWMP and IRBM monitoring stations. These sampling points provide comprehensive spatial coverage, enabling assessment of river health along the entire stretch of the Ganga as presented in Figure 1. The dataset comprises GPS-referenced physicochemical and microbial indicators: DO, pH, Electrical Conductivity, BOD, FC, and TCo, benchmarked against national standards (DO > 5 mg/L, pH 6.5–8.5, BOD < 3 mg/L, fecal coliform < 2500 MPN/100 mL). Sampling followed standardized protocols by CPCB-HQ Delhi, RD-Lucknow,

and state agencies, yielding a multivariate dataset suitable for machine-learning modeling. Spatial coverage captures ecological heterogeneity from cleaner upstream segments in Uttarakhand to pollution-intense downstream stretches in Bihar and West Bengal. This gradient provides a robust basis for predictive modeling, integrating physicochemical and microbiological indicators across state boundaries for accurate BOD prediction.

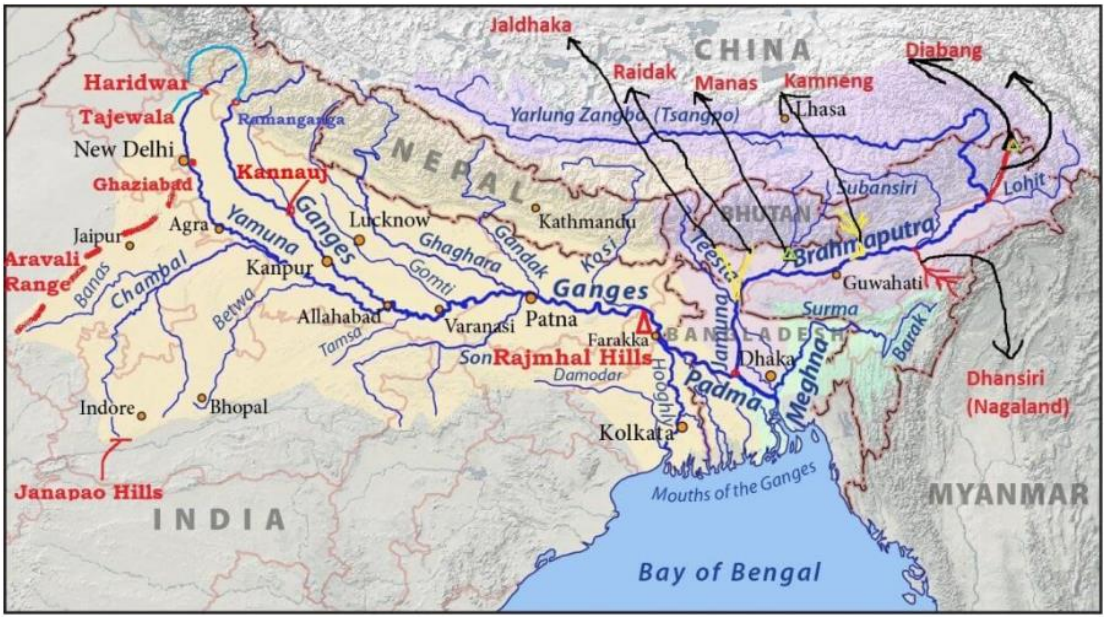

***Figure 1**: Map of the Ganga (Ganges) river*

2.2 Theory of Models

2.2.1 Fine-Tree Linear Regression (FLR)

This is known as Piecewise-Linear Regression. FLR integrates decision-tree partitioning with linear regression by fitting linear models within each tree leaf. This allows the model to capture both global nonlinearity and local linear relationships, improving predictive performance in heterogeneous datasets [14]. If the input space is partitioned into (m) disjoint regions (R1..... Rm), then

$$y\hat{} = \sum_{k=1}^{m} \mathrm{I}(\mathrm{x} \in \mathrm{Rk})\ (\beta 0k + \beta k^{T}) \quad (1)$$

where *I(.)* is the indicator function for region membership, x is the predictor vector, and *βok; βk* are the intercept and slopes for region Rk [15].

2.2.2 Interactive Linear Regression (ILR)

ILR extends standard multiple linear regression by including interaction terms between predictors. This allows modeling the combined effect of two or more variables on the response, capturing non-additive relationships [16]. The equation for two predictors is:

$$y = \beta_0 + B_1X_1 + B_2X_2 + B_{1,2}(X_1 \times X_2) + \varepsilon \quad (2)$$

where *B1,2* quantifies the interaction effect between X_1 and X_2.

2.2.3 Robust Linear Regression (RLR)

RLR reduces the influence of outliers and assumption violations in classical OLS regression. The M-estimator, introduced by Huber (1964), replaces the squared-error loss with a robust loss function to down-weight large residuals [17].

$$\min_{\beta_0 \to \beta} \sum_{i=1}^{n} \rho(\text{yi} - \beta_0 - \beta xi^{T} \beta \quad (3)$$

where ρ(.) is a robust loss function (e.g., Huber's function).

2.2.4. Stepwise Linear Regression (SWLR)

SWLR is an automated variable-selection procedure that iteratively adds or removes predictors based on statistical criteria (p-values, F-tests, or information criteria). It identifies a parsimonious subset of variables that optimally explain the dependent variable [18].

$$y\hat{} = \beta_0 + \sum_{j \epsilon S} \beta_0\hat{} X_j + \epsilon \quad (4)$$

where S is the subset of selected predictors.

2.2.5 Ensemble learning technique (ELT)

Ensemble models consist of multiple base learners whose combined predictions produce results that are typically more accurate and more stable than those of any single model. They integrate the outputs of several classifiers or predictors to improve reliability and predictive performance in both supervised and unsupervised learning tasks [19]. Previous studies also show that using two or more predictors together can significantly strengthen the forecasting ability of time-series models [20]. The literature consistently highlights that combining model outputs is an effective strategy for improving prediction efficiency in time-series applications.

2.2.6 Non-linear neural ensemble (NNE)

The NNE model consists of multiple neural network predictors that are combined through nonlinear integration to improve the performance of the learning system. This type of ensemble is widely applied in machine learning and deep learning because it strengthens model robustness and enhances predictive accuracy. In nonlinear neural ensembles, a separate neural network is trained to perform nonlinear averaging, where the outputs of the selected base models serve as inputs to the ensemble network. Each model output is assigned to a neuron in the input layer. For the FFNN-based ensemble used in this study, the tangent sigmoid activation function is applied in both the hidden and output layers, and training is carried out using the backpropagation algorithm. The optimal network structure and appropriate number of epochs are determined through a trial-and-error procedure. The nonlinear ensemble adopted here is a feedforward neural network (FFNN), as it is a widely used and well-established approach in artificial intelligence [21].

2.3 Model Preprocessing and Evaluation Measures

Before model development, the water-quality dataset consisting of 83 observations was subjected to preprocessing to ensure data integrity and suitability for regression analysis. All input variables were normalized to a [0–1] range using min-max scaling to eliminate unit-based disparities and ensure uniform influence of each predictor on model training [22]. The dataset was subsequently partitioned into training and testing subsets (70:30) for model calibration and validation [23]. Four regression models (M1–M4) were developed using different combinations of input parameters, as summarized in Table 1. Model performance was assessed using multiple statistical evaluation metrics (equations 1-4), including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Symmetric Mean Absolute Percentage Error (SMAPE) [24][25]. This provides a comprehensive measure of prediction accuracy and reliability. The methodological flowchart was presented in Figure 2.

***Table 1**: Model's Input Parameter Combination*

Models	Input Parameters Combination				
M1	DO	pH			
M2	DO	pH	Conductivity		
M3	DO	pH	Conductivity	Total Coliform	
M4	DO	pH	Conductivity	Total Coliform	Fecal Coliform

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{BOD}_{(p)} - \mathrm{BOD}_{(o)})^2 \quad (5)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{BOD}_{(p)} - \mathrm{BOD}_{(o)})^2} \quad (6)$$

$$MAE = \frac{\sum_{i=1}^{N}|\mathrm{BOD}_{(p)} - \mathrm{BOD}_{(o)}|}{N} \quad (7)$$

$$SMAPE = \frac{100}{n}\sum_{i=1}^{N}\left|\frac{|BOD'_{(o)} - \mathrm{BOD}_{(p)}|}{(|\mathrm{BOD}_{(o)}| + |BOD'_{(o)}|)/2}\right| \quad (8)$$

where BODo is the observed value, BODp is the simulated value, and BODo′ is the mean observed value.

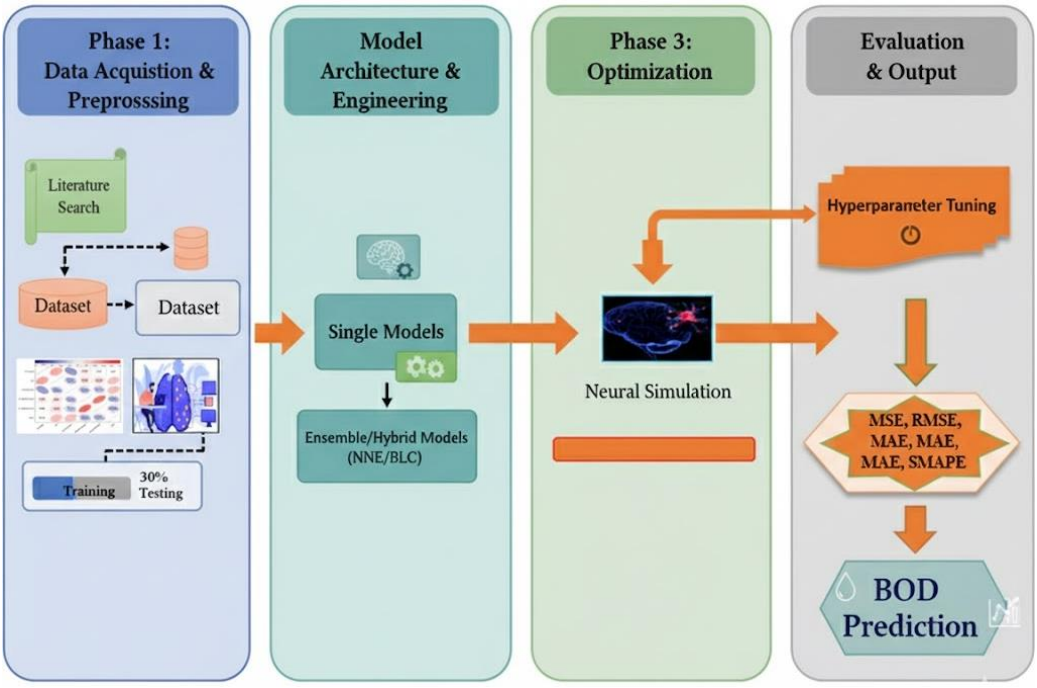

***Figure 2**: Study methodology flowchart*

3. Results and Discussion

3.1 Feature Selection and Feature Engineering

Accurate prediction of water-quality parameters requires careful selection and engineering of input features. The dataset comprised six primary indicators: BOD (mg/l), Dissolved Oxygen (mg/l), Conductivity (µmhos/cm), pH, Total Coliform (MPN/100 ml), and Fecal Coliform (MPN/100 ml). An initial exploratory data analysis, including box plots, variance assessment, and correlation analysis, was performed to identify variables with low variability, strong collinearity, or extreme outliers (see Figure 3). Domain knowledge further guided the selection of features most relevant to each model: for example, M1 focused on BOD and pH due to their direct influence on organic pollution, while M4 incorporated all six indicators to capture cumulative and synergistic effects on river water quality. Feature engineering was applied to enhance model performance and interpretability. Continuous variables were normalized to a 0–1 scale to ensure uniform contribution and mitigate bias from differing units. Interaction terms (e.g., pH vs DO) were introduced for ILR models to capture compounded environmental effects. Piecewise linear transformations in FLR models addressed local non-linear relationships, while outlier-resistant scaling in RLR minimized the influence of extreme values such as sudden spikes in coliform counts. These engineered features allowed the models to capture both linear and non-linear dynamics inherent in river water-quality parameters. This combined approach of strategic feature selection and engineering ensured that models were trained on predictors that are informative, non-redundant, and properly scaled, improving predictive accuracy, robustness, and generalizability. Beyond model performance, this methodology provides actionable insights for water resource management, as the selected and transformed features directly reflect key environmental drivers. By integrating statistical rigor with domain relevance, the study ensures that predictive models are both scientifically sound and practically applicable across the Ganga River water-quality dataset.

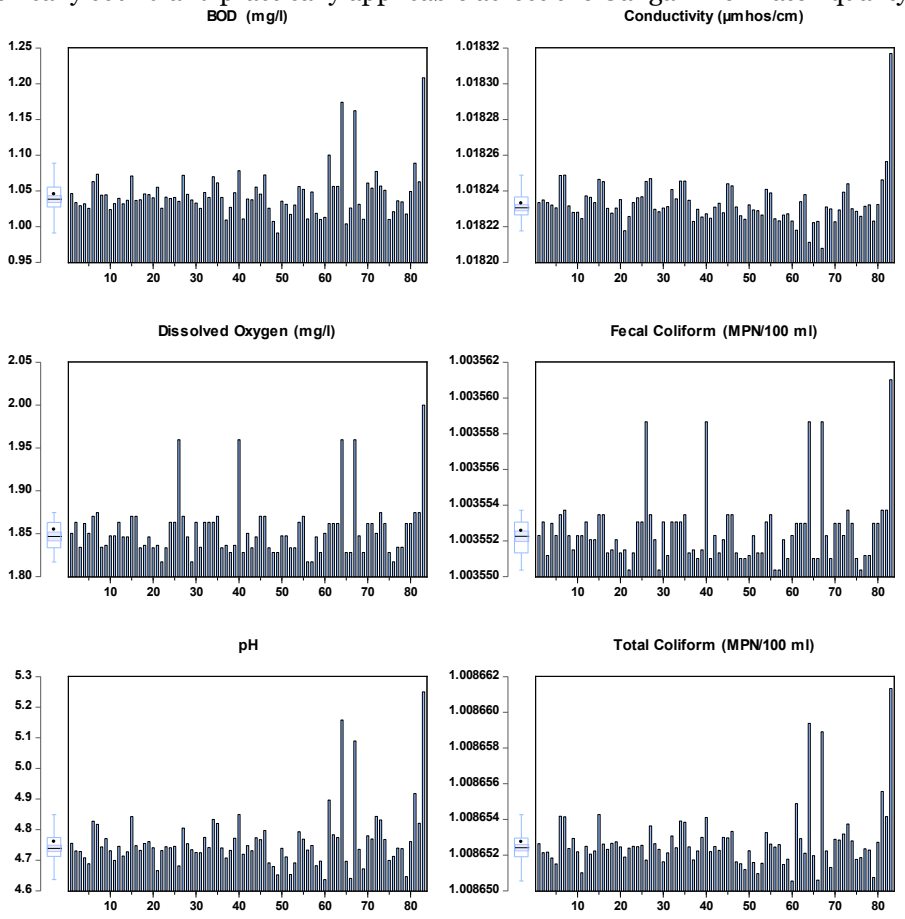

***Figure 3**: Parameter distribution pattern*

3.2 Statistical Evaluation Results

The correlation matrix elucidates the pairwise relationships among six key water-quality parameters; DO, pH, Conductivity, TC, FC, and BOD (see Figure 4). This is introduced to highlight the strength and direction of linear associations. Strong positive correlations, such as DO with FC (0.85) and DO with BOD (0.61), indicate that regions with elevated organic load also exhibit increased oxygen demand and microbial activity, suggesting that DO could serve as a proxy variable in simplified predictive models, thereby reducing dimensionality without compromising accuracy. Conversely, pH exhibits weak correlations (0.06–0.25) with most parameters, emphasizing its independent environmental influence and the necessity of retaining it in models to capture subtle but relevant effects. Weak negative associations, such as Conductivity with TC (–0.082), demonstrate minimal interdependence between ionic content and microbial counts, informing strategies to avoid multicollinearity and enhance model interpretability. These insights have direct applications in water-quality modeling: strongly correlated variables can be selectively combined to improve computational efficiency and reduce overfitting in models such as FLR and RLR, while variables with low correlations but ecological significance, like pH, enhance model generalizability. Furthermore, interaction terms between moderately correlated parameters (e.g., pH versus DO) can be incorporated into ILR frameworks to capture compounded environmental effects. Collectively, the correlation analysis provides a robust, data-driven foundation for both predictive modeling and targeted water-management interventions, ensuring that selected features are informative, non-redundant, and aligned with practical environmental objectives.

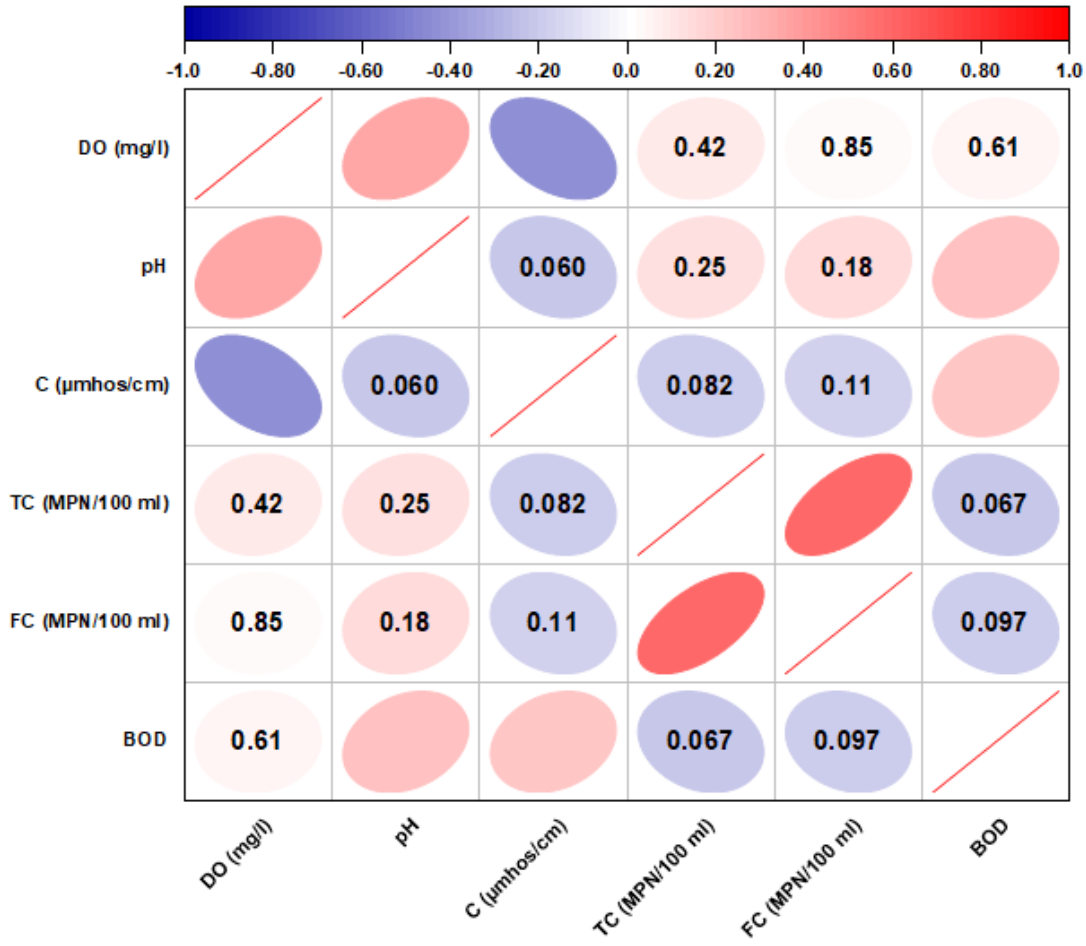

***Figure 4**: Correlation analysis matrix*

Table 2 summarizes the key descriptive statistics for six water-quality parameters used in predictive modeling: DO, pH, Conductivity, TC, FC, and BOD. The mean values indicate the central tendency of the normalized dataset, with Conductivity (0.9329) and FC (0.9040) showing relatively higher average levels compared to other parameters, suggesting dominant ionic content and microbial activity in the water samples. Standard deviations are moderate (0.145–0.217), reflecting variability across observations, while standard errors are low (0.016–0.024), indicating that the sample means are reliable estimates of the population parameters. The skewness and kurtosis values reveal the distributional characteristics of the dataset. Most parameters display negative skewness (e.g., DO: −0.267, BOD: −2.421), indicating a slight left-tail tendency, whereas pH shows near-zero skewness (0.042), reflecting approximate symmetry. High kurtosis values for Conductivity (21.73), FC (18.50), and TC (5.84) suggest heavy-tailed distributions and the presence of outliers, which should be considered in modeling to avoid bias. The range, minimum, and maximum confirm that all variables were normalized to a 0–1 scale, ensuring comparability and facilitating efficient convergence in regression models. From an application perspective, these statistics inform both model selection and feature engineering. Parameters with higher variance and extreme kurtosis may benefit from robust regression approaches (e.g., RLR) or transformation to mitigate the influence of outliers. Similarly, normalized data supports predictive models (e.g., FLR or ILR) by ensuring numerical stability and avoiding scale dominance. The descriptive analysis establishes a quantitative foundation for subsequent correlation assessment, feature selection, and predictive modeling. This verifies that each variable's statistical behavior is adequately captured and incorporated.

***Table 2**: Descriptive Statistics for Modeling Variables*

Parameter	**DO**	**pH**	**Conductivity**	**Total Coliform**	**Fecal Coliform**	**BOD**
Mean	0.6138	0.4916	0.9329	0.8342	0.9040	0.7937
Standard Error	0.0216	0.0206	0.0160	0.0239	0.0160	0.0196
Mode	0.6667	0.4188	0.9911	0.7675	0.8633	0.6875
Standard Deviation	0.1970	0.1874	0.1456	0.2175	0.1459	0.1786
Sample Variance	0.0388	0.0351	0.0212	0.0473	0.0213	0.0319
Kurtosis	0.9834	0.6279	21.7284	5.8427	18.4962	8.5913
Skewness	-0.2674	0.0415	-4.2151	-2.3894	-3.6892	2.4209
Range	1.0000	1.0188	1.0000	1.0000	1.0000	1.0000
Minimum	0.0000	0.0125	0.0000	0.0000	0.0000	0.0000
Maximum	1.0000	1.0063	1.0000	1.0000	1.0000	1.0000
Confidence Level (95%)	0.0430	0.0409	0.0318	0.0475	0.0318	0.0390

3.3 Model Performance Results

3.3.1 Single-Model Performance

Table 3 reports the predictive performance of the individual models FLR, ILR, RLR, and SWLR for both the training and testing phases. To complement these numerical results, the empirical cumulative distribution functions (CDFs) presented in Figures 5a and 5b provide additional insight into the distribution and concentration of prediction errors across the respective model structures. Among the FLR variants, FLR-4 consistently achieved the strongest predictive accuracy, reflected in a testing MSE of 0.0043 and an RMSE of 0.0657. This quantitative advantage is mirrored clearly in the CDF plots: the FLR curves, particularly those corresponding to FLR-4, show the steepest ascent and are positioned closest to the origin. This pattern indicates that many errors fall within a narrow, low-magnitude band. The close spacing between the upper and lower bounds of the CDF envelopes further suggests that FLR-based predictions are not only accurate but also stable, with limited sensitivity to noise or perturbations in the data. Such behavior underscores the strength of fuzzy regression when modelling systems are characterized by overlapping influences and nonlinear interactions, such as water-quality dynamics.

The ILR models yielded moderately higher testing errors, with ILR-4 recording MSE = 0.0090 and RMSE = 0.0950. The corresponding CDF profiles confirm this trend: the ILR curves rise more gradually than their FLR counterparts and display a broader dispersion, indicating a higher proportion of mid-range residuals. Although ILR-4 performs better than ILR-1 through ILR-3, its CDF trace still reflects a less compact error distribution, highlighting ILR's limitations in capturing subtle nonlinearities. On the other hand, RLR models demonstrated stable training behavior (MSE ≈ 0.0188–0.0190) but their performance deteriorated substantially on the testing set (MSE ≈ 0.0840–0.1018). The CDF plots illustrate this clearly; RLR curves lie noticeably to the right and exhibit the widest spread among all model classes, revealing persistent large-error contributions and a strong indication of overfitting. The divergence between training stability and testing performance implies that the robustness methods embedded in the RLR framework did not translate into improved generalization under real-world variability.

The SWLR models occupied an intermediate position. With a testing MSE of 0.0104 and RMSE of 0.1019, SWLR-4 showed modest accuracy gains over earlier SWLR variants, yet remained less precise than FLR-4. The corresponding CDF curves align with this observation, clustering more tightly than those of RLR but lacking the sharp, left-biased rise observed in the FLR family. This suggests that while stepwise feature selection mitigates some redundancy and multicollinearity, it does not fully address the nonlinear structure of the target variable. Taken together, both the statistical metrics and the CDF-based visual diagnostics converge on the same conclusion: FLR, and particularly FLR-4, provides the most accurate and reliably distributed predictions among all tested single-model approaches. The superior concentration of residuals in the FLR CDF (Figure 5) plots illustrate the model's capacity to capture uncertainty and nonlinear relationships more effectively than conventional regression frameworks. These findings reinforce the suitability of FLR as a primary modelling strategy for complex water-quality prediction tasks.

***Table 3**: Model output results using ML approaches*

Models	**TRAINING**				**TESTING**			
	MSE	RMSE	MAE	SMAPE	MSE	RMSE	MAE	SMAPE
FLR-1	0.0087	0.0935	0.0017	0.1093	0.0061	0.0782	0.0009	0.1322
FLR-2	0.0082	0.0904	0.0006	0.1056	0.0046	0.0681	0.0009	0.1199
FLR-3	0.0080	0.0897	0.0005	0.1049	0.0044	0.0664	0.0011	0.1153
FLR-4	0.0078	0.0885	0.0003	0.1022	0.0043	0.0657	0.0011	0.1150
ILR-1	0.0187	0.1369	0.0005	0.1332	0.0175	0.1324	0.0004	0.1847
ILR-2	0.0174	0.1318	0.0008	0.1274	0.0139	0.1179	0.0005	0.1567
ILR-3	0.0169	0.1302	0.0012	0.1296	0.0116	0.1076	0.0002	0.1435
ILR-4	0.0151	0.1227	0.0021	0.1179	0.0090	0.0950	0.0003	0.1412
RLR-1	0.0190	0.1379	0.0008	0.0957	0.1018	0.3190	0.0002	0.2622
RLR-2	0.0188	0.1370	0.0009	0.0966	0.0931	0.3051	0.0001	0.2529
RLR-3	0.0189	0.1376	0.0012	0.0995	0.0865	0.2942	0.0001	0.2368
RLR-4	0.0188	0.1372	0.0012	0.1002	0.0840	0.2898	0.0002	0.2354
SWLR-1	0.0187	0.1369	0.0005	0.1332	0.0175	0.1324	0.0004	0.1847
SWLR-2	0.0176	0.1326	0.0008	0.1293	0.0145	0.1204	0.0003	0.1681
SWLR-3	0.0174	0.1320	0.0012	0.1325	0.0121	0.1102	5.32E-05	0.1568
SWLR-4	0.0166	0.1290	0.0013	0.1260	0.0104	0.1019	1.71E-05	0.1479

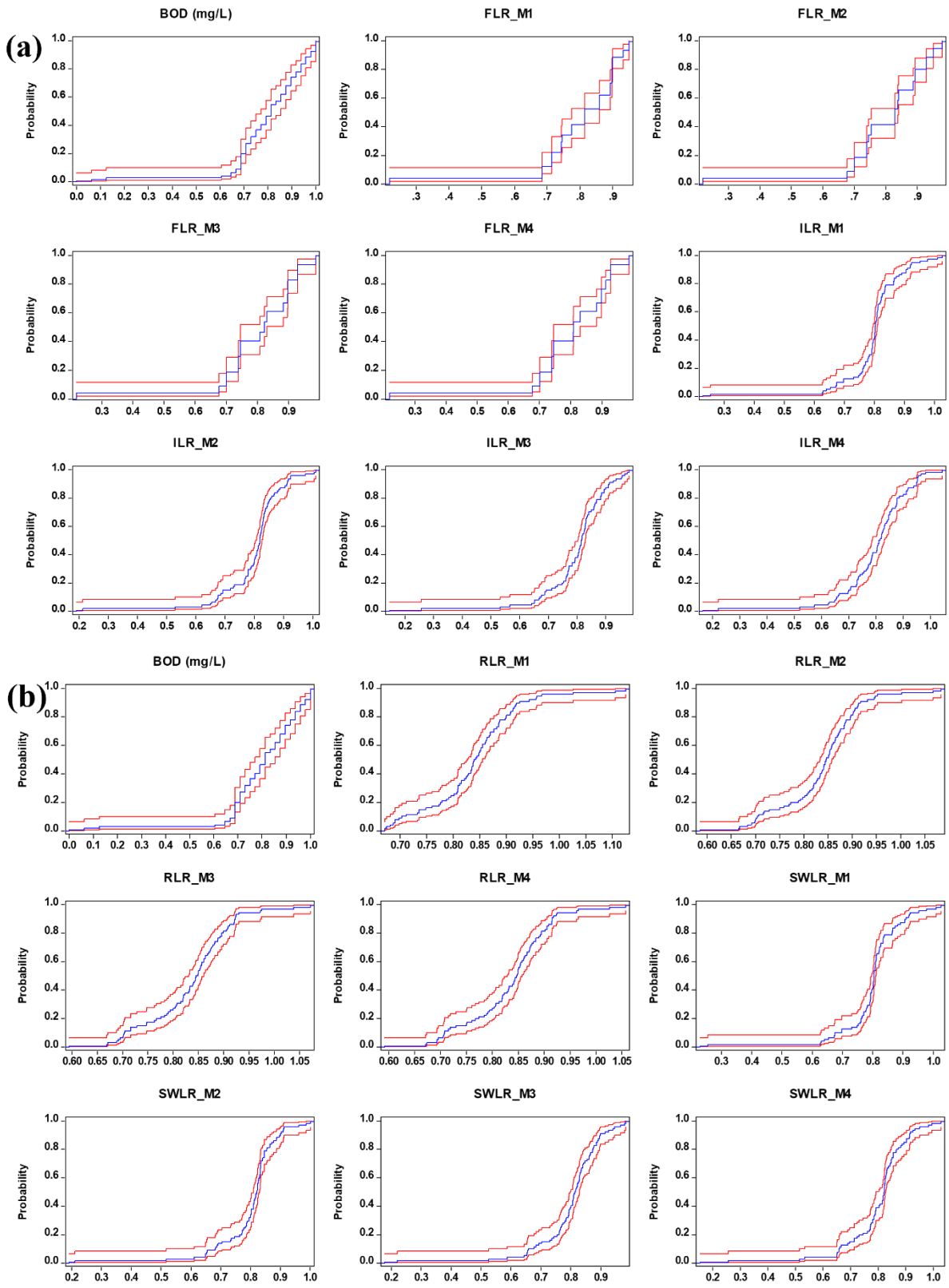

Figure 5*: Single models CDF distribution plot (a and b)*

3.3.2 Neural Network and BNE Performance

Table 4 presents the performance outcomes for the NNE and BNE frameworks built upon the four regression foundations (FLR, ILR, RLR, SWLR). Across all model families, the ensemble strategies produced a substantial enhancement in predictive accuracy relative to the corresponding single-model structures. The distributional behavior depicted in the ensemble prediction plots (Figure 6) provides a complementary perspective, illustrating how the ensembles reshape the error structure and stabilize the prediction outputs. For the FLR-based models, BNE-FLR achieved the strongest performance, with a testing MSE of 0.0031 and an RMSE of 0.0555. This represents a clear improvement over both NNE-FLR (MSE = 0.0033, RMSE = 0.0576) and the best-performing single FLR variant (FLR-4; RMSE = 0.0657). The BNE-FLR prediction plots show a notably uniform distribution, with fewer abrupt drops and tighter clustering around the upper probability band. This visual compactness indicates reduced variability and improved resilience to outlier behavior, confirming the numerical gains reported in Table 4. The ensemble gains were even more pronounced for the ILR models. BNE-ILR achieved an RMSE of 0.0392, representing nearly a sixfold improvement over the single ILR models, while NNE-ILR also produced substantial reductions in error. The associated prediction distributions reveal a striking contrast with the single ILR curves: the ensemble outputs exhibit dense, high-level plateaus with markedly fewer low-probability excursions. This indicates that the neural ensemble mechanisms successfully compensate for the linear interaction model's inherent sensitivity to noise and local gradient irregularities. A particularly notable outcome emerged from the RLR-based ensembles. While single RLR models struggled with overfitting and produced relatively high testing errors, BNE-RLR recorded a testing RMSE of 0.0381, representing one of the greatest improvements across all model families. Correspondingly, the BNE-RLR prediction plot displays a flattening of the irregular dips observed in the single RLR outputs, with the boosted ensemble effectively suppressing error spikes and stabilizing the prediction profile. This confirms that boosting can counteract the brittleness and variance amplification typically associated with robust regression in small or noisy datasets.

SWLR ensembles also benefited from the neural aggregation strategies. BNE-SWLR achieved an RMSE of 0.0398, a large improvement over the best single SWLR variant (RMSE = 0.1019). In the ensemble plots, the SWLR-based distributions show a clear upward shift with fewer pronounced downward deviations, revealing that the ensemble successfully mitigates the instability introduced by stepwise variable selection. The resulting curves resemble those of the FLR-based ensembles in terms of smoothness and compactness, despite SWLR's more rigid modelling structure. Taken together, the numerical results and plot-based diagnostics converge on the same conclusion: neural ensembles, especially boosted ensembles, significantly enhance predictive reliability across all regression families. The ensembles reduce both systematic errors (lower bias) and random fluctuations (lower variance), while producing prediction distributions that are smoother, more concentrated, and far less prone to extreme deviations. These attributes confirm the ensembles' superior ability to generalize across varying water-quality conditions and highlight the value of integrating neural aggregation mechanisms with traditional regression frameworks.

***Table 4**: Neural Network Ensemble Performance Metrics*

Models	**Training**				**Testing**			
	MSE	**RMSE**	**MAE**	**SMAPE**	**MSE**	**RMSE**	**MAE**	**SMAPE**
NNE-FLR	0.0062	0.0786	0.0002	0.0859	0.0033	0.0576	5E-05	0.1078
BNE-FLR	0.0061	0.0780	7.97E-05	0.0826	0.0031	0.0555	1.27E-05	0.1009
NNE-ILR	0.0059	0.0768	0.0016	0.1048	0.0030	0.0544	0.0002	0.0543
BNE-ILR	0.0034	0.0587	0.0018	0.0817	0.0015	0.0392	0.0007	0.0292
NNE-RLR	0.0059	0.0767	0.0002	0.0804	0.0039	0.0622	0.0014	0.0588
BNE-RLR	0.0013	0.0366	0.0012	-0.0033	0.0015	0.0381	0.0003	0.0376
NNE-SWLR	0.0046	0.0676	0.0020	0.0960	0.0020	0.0449	0.0008	0.0424
BNE-SWLR	0.0039	0.0625	0.0003	0.0135	0.0016	0.0398	0.0002	0.0316

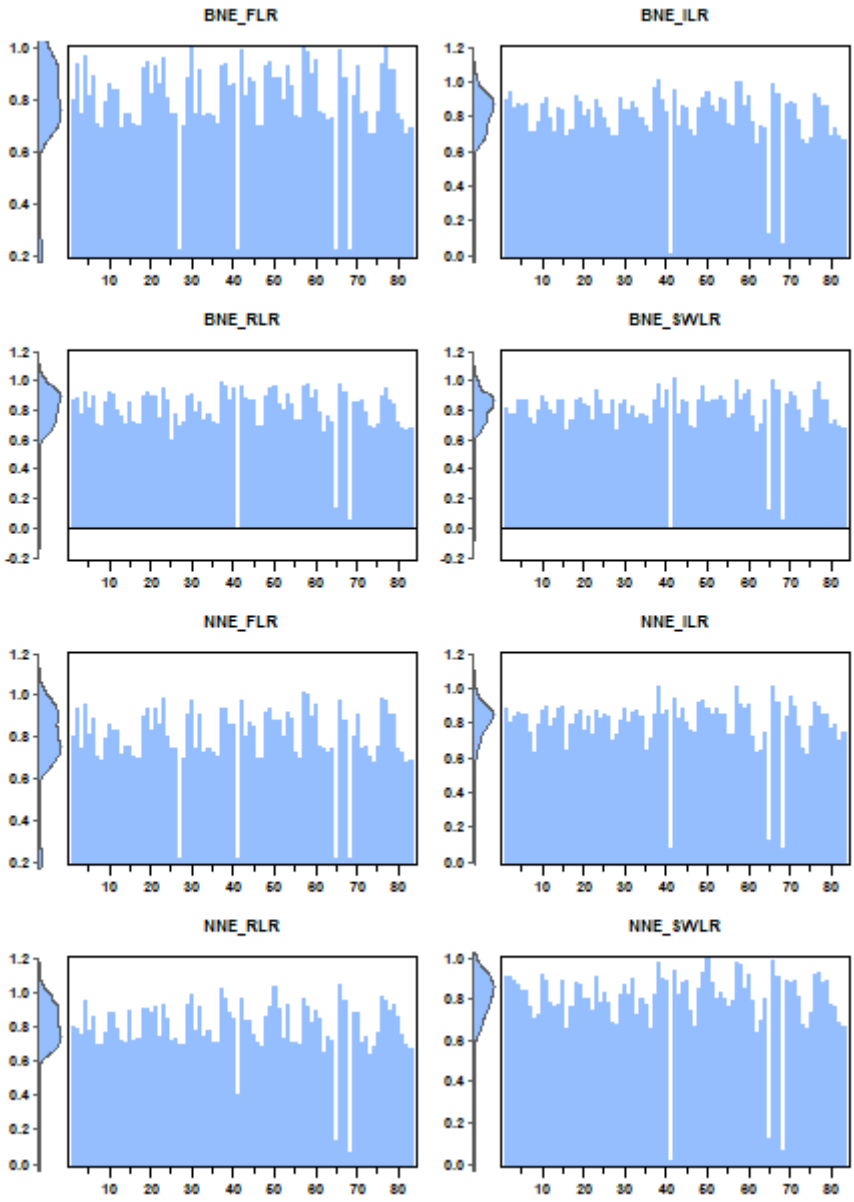

***Figure 6**: Ensemble results distributional behavior*

3.4 Real World Applications of Predictive Models for BOD in the Ganga River

The radar plots (Figure 7) presented earlier compare the predictive performance (RMSE, MAE, SMAPE) across various model types, including regression-based models (e.g., FLR, ILR, RLR, SWLR) and more complex neural–network–based ensemble models. In many practical applications, simpler regression models (e.g., FLR) tend to achieve moderate accuracy; their lower RMSE and MAE relative to some more naive regression variants, suggest that they can capture some of the nonlinear patterns typical in river water quality dynamics. Nevertheless, such models are limited by inherent model biases, restricted functional form, and potential overfitting or underfitting, especially when river chemistry responds to complex, interacting environmental drivers. By contrast, empirical evidence from recent water quality modelling studies confirms that neural network and ensemble approaches, especially hybrid or stacked models, often deliver substantially improved predictive accuracy. For example, in a survey on the Karun River, a hybrid model combining wavelet-transformed features with a Random Forest (RF) base achieved low error values for BOD prediction, outperforming both pure tree-based and regression tree models. [26] Similarly, an investigation on the Jinjiang River basin proposed a hybrid model based on a Long Short-Term Memory (LSTM) neural network with discrete wavelet transform (DWT)

preprocessing (ANN WT LSTM), which delivered superior performance compared with conventional models. A more recent demonstration for the Godavari River Basin used a stacking ANN meta model built on multiple machine learning base learners (e.g., RF, boosting methods) and showed that such ensemble models significantly increased the coefficient of determination for BOD predictions, as proved by [27]. These findings support the inference from the radar plots that bilayer neural network ensembles can robustly capture complex, nonlinear, and interacting influences on water quality than simpler models [28].

From a management perspective, high-accuracy predictive models offer tangible benefits for real-world water quality control. First, early warning systems may be built to provide timely and reliable forecasts of BOD spikes or organic load surges that can trigger alerts before oxygen depletion events threaten aquatic life or public health. Second, treatment optimization becomes feasible when wastewater treatment plants or remediation systems can adjust aeration, chemical dosing, or discharge control based on predicted BOD fluctuations rather than relying solely on measured BOD (which requires a lagging incubation period). Third, pollution source identification and hotspot detection can be supported by coupling spatially distributed water quality data (e.g., from multiple monitoring stations) with model predictions, so that stakeholders can map where organic loading is most severe, thereby aiding targeted regulatory action or cleanup. Fourth, predictive models offer decision support for river restoration: what-if scenarios (e.g., changes in land use, discharge regulations, runoff controls) can be simulated to evaluate their potential impact on BOD and overall water quality. Finally, when integrated with real-time or near real-time monitoring networks (sensors for DO, TSS, flow, temperature, etc.), these models can power dynamic, adaptive management: as new data come in, forecasts update, enabling real-time response to pollution events. The empirical literature supports the conclusion that a hybrid framework combining multivariate regression (for interpretability) with neural network or ensemble models (for predictive power) yields a robust, scalable, and practical tool for real-time water quality prediction and management. This strongly suggests that applying BNE style models to a complex, large river system such as the Ganga River is not only methodologically defensible but potentially transformative for ecological integrity and public health protection in practice.

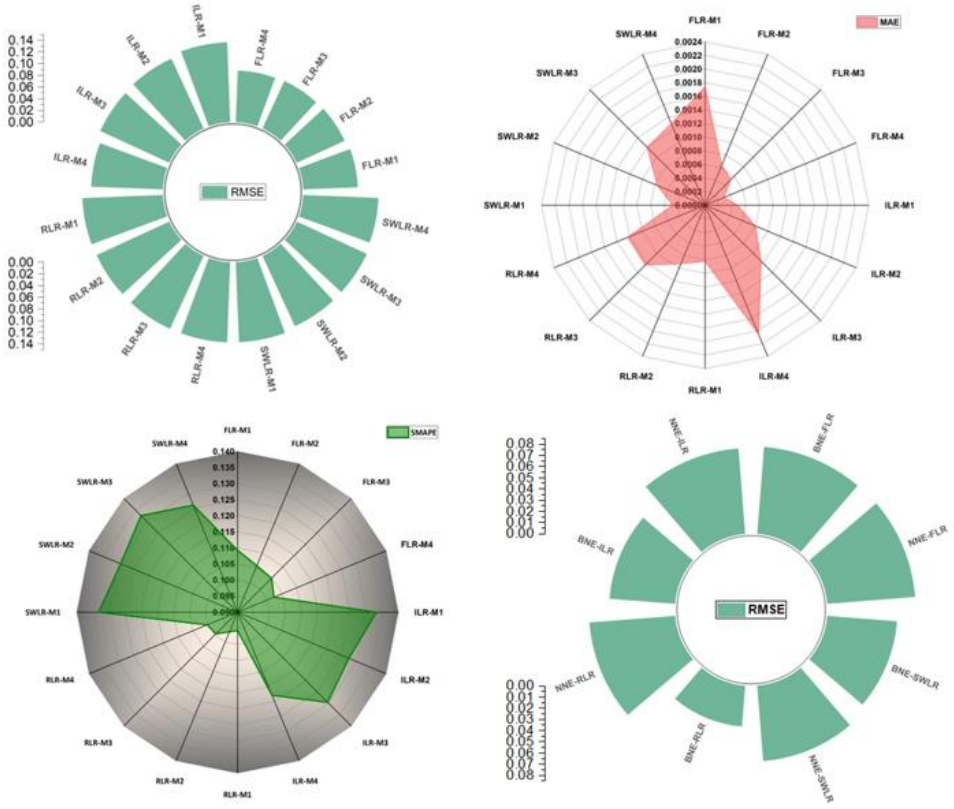

***Figure 7**: Spider and radar plots for both single and ensemble models*

4. Conclusion

This study demonstrates the efficacy of a hybrid data-driven framework combining multivariate regression models with neural network ensemble techniques for predicting Biochemical Oxygen Demand in the Ganga River. Among the regression models, Fine-Tree Linear Regression exhibited superior performance relative to Interactive Linear Regression, Robust Linear Regression, and Stepwise Linear Regression, with FLR-4 achieving the lowest testing errors. Nonetheless, the integration of regression outputs into Bilayer Neural Ensemble models substantially enhanced predictive accuracy and stability, with BNE-RLR and BNE-ILR producing the most reliable forecasts across all evaluation metrics (MSE, RMSE, MAE, and SMAPE). Correlation analysis highlighted that physiochemical parameters such as pH, conductivity, and dissolved oxygen contributed moderately to BOD variation, while microbial indicators showed weak associations, emphasizing the limitations of linear approaches in capturing complex, nonlinear interactions inherent in river water-quality dynamics. The neural network ensemble framework effectively addressed these limitations, leveraging complementary strengths of individual regression models to reduce systematic and random errors. Conclusively, the findings indicate that coupling multivariate regression with neural network ensemble modeling offers a robust, scalable, and practical tool for real-time and accurate BOD prediction. This approach not only improves forecasting reliability but also provides actionable insights for water-quality management, early warning systems, pollution source identification, and informed decision-making for ecological restoration in major river systems such as the Ganga. The study underscores the potential of hybrid predictive frameworks to enhance environmental monitoring and support sustainable riverine management practices.

Competing Interests: The authors declare that they have no competing interests.

Data Availability Statement: The supported data associated with this researcher is available upon request from the corresponding author.

References

[1] S. Srivastav, K. Guleria, and S. Sharma, “Waste Segregation using Deep Learning-based Convolutional Neural Network Model,” in 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), 2023, pp. 1–6. doi: 10.1109/GCAT59970.2023.10353523.
[2] A. Kumari, N. S. Maurya, and B. Tiwari, “Hospital wastewater treatment scenario around the globe,” in Current developments in Biotechnology and Bioengineering, Elsevier, 2020, pp. 549–570.
[3] WHO, Global analysis of healthcare waste in the context of COVID-19: status, impacts and recommendations. World Health Organization, 2022.
[4] O. S. Kushwaha, H. Uthayakumar, and K. Kumaresan, “Modeling of carbon dioxide fixation by microalgae using hybrid artificial intelligence (AI) and fuzzy logic (FL) methods and optimization by genetic algorithm (GA),” Environ. Sci. Pollut. Res., vol. 30, no. 10, pp. 24927–24948, 2023.
[5] R. M. Adnan et al., “Modelling biochemical oxygen demand using improved neuro-fuzzy approach by marine predators algorithm,” Environ. Sci. Pollut. Res., vol. 30, no. 41, pp. 94312–94333, 2023.
[6] T. Li, T. Lan, H. Zhang, J. Sun, C.-Y. Xu, and Y. D. Chen, “Identifying the possible driving mechanisms in Precipitation-Runoff relationships with nonstationary and nonlinear theory approaches,” J. Hydrol., vol. 639, p. 131535, 2024.
[7] I. I. Ismail et al., “Ensemble Machine Learning Technique Based on Gaussian Algorithm for Stream Flow Modelling,” Techno-computing J., vol. 1, no. 2, pp. 1–17, 2025.
[8] Z. Zhao, L. Alzubaidi, J. Zhang, Y. Duan, and Y. Gu, “A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations,” Expert Syst. Appl., vol. 242, p. 122807, 2024, doi: https://doi.org/10.1016/j.eswa.2023.122807.
[9] A. Montusiewicz, Methods for Enhancing Biogas Production. Routledge, 2024.
[10] R. Choudhary, A. Kumar, P. C, M. M. Naik, M. Choudhury, and N. A. Khan, “Predicting water quality index using stacked ensemble regression and SHAP based explainable artificial intelligence,” Sci. Rep., vol. 15, no. 1, p. 31139, 2025.
[11] A. Aldrees, A. M. Jibrin, S. Dan’azumi, I. A. Mahmoud, U. U. Aliyu, and S. I. Abba, “Explainable machine learning framework for assessing groundwater quality and trace element contamination in Eastern Saudi Arabia,” Sci. Rep., 2025.
[12] M. K. Mostafa, A. S. Mahmoud, M. S. Mahmoud, and M. Nasr, “Computational-based approaches for predicting biochemical oxygen demand (BOD) removal in adsorption process,” Adsorpt. Sci. Technol., vol. 2022, p. 9739915, 2022.

[13] R. Mishra, R. Singh, and C. B. Majumder, "Forecasting biochemical oxygen demand (BOD) in River Ganga: a case study employing supervised machine learning and ANN techniques," Sustain. Water Resour. Manag., vol. 11, no. 1, p. 9, 2025.
[14] A. Bemporad, "A piecewise linear regression and classification algorithm with application to learning and model predictive control of hybrid systems," IEEE Trans. Automat. Contr., vol. 68, no. 6, pp. 3194–3209, 2022.
[15] N. D. Vanli and S. S. Kozat, "A comprehensive approach to universal piecewise nonlinear regression based on trees," IEEE Trans. Signal Process., vol. 62, no. 20, pp. 5471–5486, 2014.
[16] J. Fox, An R and S-Plus companion to applied regression. Sage, 2002.
[17] P. J. Huber, "Robust estimation of a location parameter," in Breakthroughs in statistics: Methodology and distribution, Springer, 1992, pp. 492–518.
[18] L. L. Nathans, F. L. Oswald, and K. Nimon, "Interpreting multiple linear regression: a guidebook of variable importance.," Pract. assessment, Res. Eval., vol. 17, no. 9, p. n9, 2012.
[19] S. I. Abba et al., "Hybrid machine learning ensemble techniques for modeling dissolved oxygen concentration," IEEE Access, vol. 8, pp. 157218–157237, 2020, doi: 10.1109/ACCESS.2020.3017743.
[20] M. Sakib, S. Mustajab, and M. Alam, "Ensemble deep learning techniques for time series analysis: a comprehensive review, applications, open issues, challenges, and future directions," Cluster Comput., vol. 28, no. 1, p. 73, 2025.
[21] S. I. Abba, V. Nourani, and G. Elkiran, "Multi-parametric modeling of water treatment plant using AI-based non-linear ensemble," J. Water Supply Res. Technol., vol. 68, no. 7, pp. 547–561, Nov. 2019, doi: 10.2166/AQUA.2019.078.
[22] A. M. Jibrin et al., "Tracking the impact of heavy metals on human health and ecological environments in complex coastal aquifers using improved machine learning optimization," Environ. Sci. Pollut. Res., vol. 31, no. 40, pp. 53219–53236, 2024.
[23] U. U. Aliyu et al., "Biomass and Bioenergy Optimizing biomedical waste generation modeling using quantum machine learning and economic development indicators," vol. 204, no. July 2025, 2025.
[24] M. A. Sanjrani, X. Gang, and S. N. A. Mirza, "A review on textile solid waste management: Disposal and recycling," Waste Manag. Res., Jul. 2024, doi: 10.1177/0734242X241257093/ASSET/IMAGES/LARGE/10.1177_0734242X241257093-IMG2.JPEG.
[25] I. A. Mahmoud, U. J. Muhammad, S. J. Kawu, and M. Mukhtar, "Enhancing Energy Demand Prediction Using Elman Neural Network and Case Study Enhancing Energy Demand Prediction Using Elman Neural Network and Support Vector Machine Model : A Case Study in Lagos State , Nigeria," no. November, 2024, doi: 10.37256/aie.5220244396.
[26] S. I. Abba, S. J. Hadi, and J. Abdullahi, "River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques," Procedia Comput. Sci., vol. 120, pp. 75–82, 2017.
[27] J. Wu and Z. Wang, "A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory," Water, vol. 14, no. 4, p. 610, 2022.
[28] M. A. Novianta, B. Warsito, and S. Rachmawati, "Monitoring river water quality through predictive modeling using artificial neural networks backpropagation.," AIMS Environ. Sci., vol. 11, no. 4, 2024.