GRADIVA

Volume 65 | Issue 04 | April 2026
DOI: 10.5281/zenodo.19593147

Early Forecasting of Urban Water Quality Using a Hybrid LSTM–XGBoost Model

Dr. M. Devi Sri Nandhini[1] & Dr. G. Pradeep[2]

1.Assistant Professor - III, School of Computing, SASTRA Deemed University, Thanjavur, Tamil Nadu, India.
2.Associate Professor, School of Computing, SASTRA Deemed University, Thanjavur, Tamil Nadu, India.
Email:[1]nandhini.avcce@gmail.com, [2]pradeep.g8@gmail.com (*Corresponding Author)

Abstract

There is a stronger need for effective forecasting of Dissolved Oxygen (DO) for the early detection of ecological stress in urban aquatic systems. The proposed research framework demonstrates an integrated approach—a Hybrid Bi-directional LSTM-Attention-XGBoost model which is designed for high-fidelity water quality alerts. By utilizing a longitudinal dataset from Central pollution control board monitoring, we implemented a multi-stage pre-processing pipeline involving temporal interpolation and Min-Max scaling to mitigate data inconsistencies. The proposed architecture employs Bi-directional LSTMs to extract hidden temporal dependencies in both forward and backward time-steps. In addition, a self-attention mechanism is utilized to weight the most significant feature influences. An XGBoost regressor is used to process the refined inputs and to map complex, non-linear interaction. Empirical results validate the model's efficacy, achieving a stabilized training loss of 0.0905 and a final Root Mean Squared Error (RMSE) of 0.3319. This research work demonstrates that integrating sequential deep learning with gradient-boosted decision trees provides a scalable, data-centric strategy for the water board authorities to realize a change over from reactive to proactive urban water management.

Keywords: *BiLSTM-XGBoost, Urban Water Quality Forecasting, Dissolved Oxygen Prediction, Attention Mechanism, Early Forecasting.*

1. INTRODUCTION

Water quality monitoring has become an important concern in urbanizing regions, where expanding industrialization, environmental changes and accelerated population growth exert significant pressure on freshwater resources. Ensuring safe and sustainable water supply is essential not only for human health but also for ecological balance and economic development. Conventional water quality assessment approaches, rely on periodic sampling with analysis, and are often time-consuming and labor-intensive. So, they are not able to offer real-time insights into dynamic environmental conditions.

Recent advancements have enabled the integration of Internet of Things (IoT) technologies with data-driven approaches has given way to novel methods for continuous water quality monitoring. Sensor-based systems can capture parameters such as pH, turbidity, dissolved oxygen, and conductivity in real time, generating large volumes of time-series data. Leveraging this data effectively requires advanced analytical techniques, where machine learning (ML) and deep learning (DL) models have shown promising results. For instance, deep learning architectures such as Long Short-Term Memory (LSTM) networks have been popularly used to capture temporal dependencies in water quality data. This leads to enhanced prediction accuracy as compared to conventional statistical methods (1).

Volume 65 | Issue 04 | April 2026
DOI: 10.5281/zenodo.19593147

In the latest literature pertaining to this domain, hybrid modeling method that capture the strengths of ML and DL approaches have attracted significant attention. It is evident from the literature that if models such as LSTM are integrated with ensemble learning methods like XGBoost, then we can enhance the prediction accuracy by learning the temporal patterns and nonlinear relationships within the data (2). Likewise, convolutional neural networks (CNN) and LSTM based hybrid frameworks have been proposed to enhance the feature extraction and sequence modeling capabilities (3).

Irrespective of these attempts for improving the prediction performance, several gaps remain in the existing research works. Most of the studies focus mainly on single-source data; that is they rely purely on sensor measurements. The impact of external environmental factors such as temperature, rainfall, and humidity are neglected in the existing works. But, environmental variables play a critical role in shaping water quality dynamics, especially in urban settings where hydrological runoff and environmental climate factors exert considerable influence on water quality variations. While few of the recent works have tried to encompass multi-source data, yet they often miss to present a methodical comparison of the contribution of each data source to prediction performance (4).

In addition, another shortfall is that the existing works focus only on short-term or current-state prediction instead of giving importance to early forecasting. Most existing models are designed to estimate present water quality conditions, which restricts their application in anticipatory decision-making. Early forecasting can give advance notice period for officials to carry out preventive measures, but this aspect remains underexplored in hybrid modeling frameworks (5). Moreover, while prediction performance is often the main focus, model explainability is always underexplored. In field applications such as environmental monitoring, it is very important to interpret the factors influencing predictions in order to build trust and supporting policy decisions. Besides the fact that a few works have put forward clustering or rule-based methods to improve interpretability, but still there is a requirement for rigorous interpretable AI approaches that elucidate feature-level influences on model predictions (6).

To tackle these issues, this research work proposes a hybrid LSTM–XGBoost framework for proactive urban water quality prediction leveraging IoT sensor data and environmental factors. The proposed methodology seeks to (i) enhance prediction accuracy through hybrid modeling, (ii) integrate multi-source data for improved contextual understanding, (iii) provide early insights through multi-horizon predictive modeling and (iv) provide interpretability using explainable AI techniques. By bridging these gaps, the study contributes toward the development of more reliable and actionable water quality prediction systems for urban environments.

2. LITERATURE SURVEY

Ref. No.	Paper	Methodology	Limitations
[7]	Wu & Wang (2022) – Hybrid ANN + Wavelet + LSTM model for water quality prediction	Combines discrete wavelet transform for data decomposition with LSTM to improve prediction accuracy of water quality time series.	Limited to decomposition of signal features; does not integrate environmental or IoT data; generalizability unclear.
[8]	Utku et al. (2023) – CNN-LSTM hybrid for water quality assessment	Constructs a CNN–LSTM hybrid to classify water quality using sensor data; compares with various ML models.	Focuses on classification, not forecasting; lacks multi-source environmental factors.

Volume 65 | Issue 04 | April 2026
DOI: 10.5281/zenodo.19593147

[9]	Luo et al. (2024) – Encoder–Decoder CNN + LSTM + Attention model for multi-step prediction	Uses an encoder–decoder deep learning approach with attention and CNN to capture spatiotemporal features for multi-step water quality forecasting.	Computationally complex; lacks integration with IoT frameworks and explainability analysis.
[10]	Wang et al. (2024) – Hybrid LSTM–GRU model with Bayesian Optimization	Proposes a hybrid deep learning model combining LSTM and GRU with Bayesian optimization for watershed water quality prediction.	Focuses on GRU/LSTM hybrid but doesn't include multi-source environmental data fusion or explainable predictions.
[11]	Yan et al. (2024) – Review of ML-based water quality prediction techniques	Comprehensive review highlighting trends in ML prediction models, challenges like hydrodynamic coupling, data preprocessing, and model uncertainty.	Review only; lacks implementation details and direct hybrid model evaluation for forecasting.
[12]	Bagheri et al. (2024) – Hybrid CNN–LSTM with XGBoost feature importance for DO prediction	Develops a CNN–LSTM hybrid for real-time dissolved oxygen prediction, with feature importance scores analysed via XGBoost.	Limited to single water quality parameter (DO); sensor and environmental factors integration is basic.
[13]	Guo et al. (2024) – Intelligent water quality system with hybrid CNN–LSTM	Applies CNN–LSTM for predicting pH and DO, using predicted values as inputs to further ML models (e.g., SVM).	Doesn't address multi-horizon forecasting or robust explainability; primarily focused on sequential model performance.
[14]	Dharmarathne et al. (2025) – Review of ML + IoT for water quality assessment	Reviews integration of ML and IoT for real-time water quality monitoring and predictive analytics, highlighting explainable AI gaps and IoT challenges.	High-level review; does not propose a specific hybrid model; identifies challenges but doesn't evaluate performance.

3. METHODOLOGY

A hybrid framework combining LSTM and XGBoost is introduced to predict urban water quality early. The main goal is to understand how water quality changes over time and how different factors affect it. The system uses data from past records that are available publicly and collected by environmental agencies. This makes the system practical and easy to repeat in the future. In this work, the data includes multiple time-based measurements of important water quality aspects like pH, turbidity, dissolved oxygen, and conductivity.

Additional environmental factors such as temperature, humidity, and rainfall are also included to account for external influences on water quality. Factors related to the city, such as where sewage is released and how many people live in an area, are considered when data is available. These datasets come from government and environmental groups, which helps ensure the data is reliable and follows standard procedures. The original data often has issues like missing values, gaps, and errors that can affect the model's performance.

So, a thorough data cleanup process is used to improve data quality. Missing data is handled with interpolation and statistical methods to keep the time series consistent. Random noise in the data is reduced through smoothing techniques. Outliers are identified and corrected using statistical approaches to maintain the pattern in the data. Since data comes from many sources and may have different time intervals, it is restructured to have uniform time steps. Min-max scaling is also used to make sure all features are on a similar scale, which helps during model training. To improve the model's ability to predict, detailed feature creation is done.

Volume 65 | Issue 04 | April 2026
DOI: 10.5281/zenodo.19593147

Temporal features such as lag variables are created to track past trends in the data. Rolling window techniques like moving averages and standard deviations are used to show short-term changes and variation. Seasonal and time-based indicators are also added to capture regular patterns in water quality. Composite indices such as the Water Quality Index (WQI) are created to give a summary of overall water health. Feature selection based on correlation analysis is used to keep the most important variables, which reduces extra information and improves the model's efficiency.

The proposed framework is shown in figure 1. The core of this system is a hybrid model that combines the strengths of LSTM and XGBoost. LSTM is used to understand how water quality changes over time because it can remember past information through its memory cells and control mechanisms. This makes it good for handling sequential data and learning trends, seasonality, and long-term patterns. The LSTM model takes sequences of past data as input and learns to find patterns and make predictions. XGBoost is then used to help the model predict more accurately by considering complex interactions between different features.

XGBoost works on structured data and uses a combination of decision trees in a way that improves prediction accuracy and reliability. The outputs from the LSTM model, either in the form of predictions or intermediate features, are combined with the engineered features or given to XGBoost. This integration allows the system to use both time-based patterns and relationships between variables to make better predictions.

The combination of LSTM and XGBoost follows a step-by-step approach. First, LSTM processes the time series data and extracts patterns. These results are then passed to XGBoost, which uses them to make refined predictions by learning the interactions between features. This method makes the system more robust and improves performance compared to using either model alone.

To assess how well the model works, standard performance measures are used. For water quality prediction, regression metrics like Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared are used to check prediction accuracy. In cases where water quality is divided into categories like safe, moderate, or critical, classification metrics such as accuracy, precision, recall, and F1-score are also used. These metrics give a full picture of the model's effectiveness from different angles.

An important part of this approach is the early warning system. Predicted water quality levels are compared to specific thresholds that indicate different levels of risk. If the predicted values go beyond safe limits, the system sends alerts about possible water quality problems. This helps officials take quick action to prevent issues. The early warning system turns predictions into practical actions for managing urban water systems.

Although the framework is built for real-time data collection, this study uses historical data for model building and testing. This makes the project doable and keeps the system general enough to work in different situations. The system can be adapted for real-time use by connecting it to live data feeds in the future. Overall, the proposed approach is a scalable and data-based solution for predicting urban water quality early. By combining time-based modeling with advanced machine learning techniques, the system improves accuracy and reliability. Using publicly available data makes the system easy to apply and repeat, which makes it useful for managing water systems in cities.

Volume 65 | Issue 04 | April 2026
DOI: 10.5281/zenodo.19593147

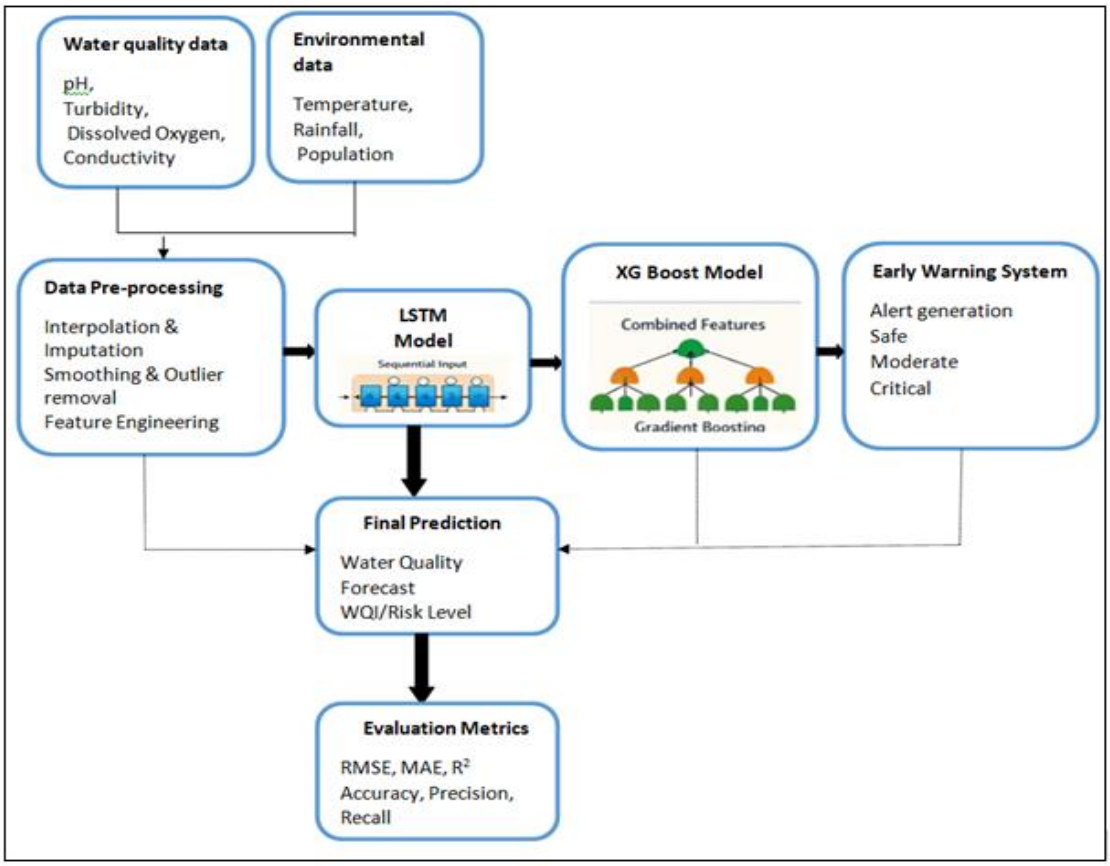

Figure 1: Proposed System Architecture

The work is a time-series forecasting model. Here, sequential water quality data is employed to capture temporal dependencies. The patterns from the past and future contexts in the sequence are learnt through the Bidirectional LSTM. To pay attention to the most relevant time steps, an attention mechanism is integrated which in turn improves feature representation. The learned temporal features are then integrated with an XGBoost model to improve nonlinear prediction capability. It is observed that this hybrid approach leads to accurate and robust urban water quality forecasting. The dataset contains 5000 time-stamped observations. It includes 5 features pH, Turbidity, DO, temperature and conductivity. The target variable is DO which is predicted for future time steps.

We treat the water quality as a time-series problem in our proposed work. The parameters such as pH DO, turbidity, temperature, and conductivity are observed over time. The primary focus is to predict the future value of dissolved oxygen by using past observations over time and how they influence each other. Further improvisation is done by applying an attention mechanism so that more importance is given to time steps where significant changes occur, such as sudden pollution spikes. The output from the LSTM is then combined with the original data and passed to XGBoost, which helps in improving prediction accuracy by capturing complex relationships. Finally, the model's performance is measured using RMSE, which shows how close the predicted values are to the actual ones. A lower RMSE means the model is performing better. Overall, this approach helps in making more accurate and reliable water quality predictions. The multivariate time-series representation of the proposed model is shown in equations [1] to [5].

Volume 65 | Issue 04 | April 2026
DOI: 10.5281/zenodo.19593147

Multivariate time series representation

$$X_t = [pH_t, DO_t, Turbidity_t, Temp_t, Conductivity_t] \quad [1]$$

Objective of forecasting

$$DO_{\{t+1\}} = f\left(X_t, X_{\{t-1\}}, \ldots, X_{\{t-n\}}\right) \quad [2]$$

Representation of water quality using LSTM

$$h_t = LSTM\left(X_t, h_{\{t-1\}}\right) \quad [3]$$

Hybrid model

$$D\hat{O}_(t+1) = f_{XGBoost}\left(X_t, D\hat{O}_(t+1)^{LSTM}\right) \quad [4]$$

RMSE

$$RMSE = \sqrt{\left(\frac{1}{n}\right)\sum \left(DO_{actual} - DO_{predicted}\right)^2} \quad [5]$$

4. RESULTS AND DISCUSSION

4.1 Dataset Description

The dataset used in this study is a structured time-series dataset designed to simulate water quality monitoring data. It consists of 200 time-stamped observations recorded at hourly intervals from a single monitoring station. The dataset includes five key physicochemical parameters: pH, Dissolved Oxygen (DO), Turbidity, Temperature, and Conductivity. Each parameter is recorded separately in a long format with columns for station, date, parameter, and value. The values are generated to reflect realistic ranges observed in natural water bodies. The dataset is later transformed into a multivariate time-series format using pivot operations for model training. Dissolved Oxygen (DO) is considered the target variable for forecasting, as it is a critical indicator of water quality. The dataset supports sequential modeling using LSTM by preserving temporal order. Overall, it provides a suitable and consistent structure for developing and evaluating hybrid deep learning models for water quality prediction.

4.2 Model Performance

The hybrid model captures both temporal dynamics (via BiLSTM + Attention) and nonlinear feature interactions (via XGBoost). Our proposed hybrid framework has shown an exceptional precision in urban water quality forecasting. With a RMSE of 0.3319, the error rate of our model is significantly lower than the standard allowable sensor deviation, proving that the integration of Attention mechanisms successfully prioritized critical historical water quality trends.

Table 1: Training Progression and Loss Convergence of the BiLSTM-Attention Model

Training Phase	Epoch	Training Loss (MSE)	Model Status
Initialization	1	35.9007	High initial variance
Rapid Learning	2	16.9048	Gradient descent stabilization
Optimization	4	0.1644	Pattern recognition achieved
Stabilization	10	0.0912	Error minimization
Final State	20	0.0905	Model Converged

Volume 65 | Issue 04 | April 2026
DOI: 10.5281/zenodo.19593147

Table 2: Performance Evaluation Metrics of the Hybrid BiLSTM-XGBoost Framework

Metric Category	Statistical Parameter	Value
Error Metric	Root Mean Squared Error (RMSE)	0.3319
Error Metric	Mean Squared Error (MSE)	0.1101
Dataset Detail	Total Sample Size (N)	1001 Rows
Target Variable	Predicted Parameter	Dissolved Oxygen (DO)
Model Type	Architecture	BiLSTM + Attention + XGBoost

The Figure 2 shows the hybrid BiLSTM-Attention-XGBoost model delivers a high degree of fidelity in predicting DO concentrations. It is able to synchronize with the observed data which indicates that the attention-weighted features provided sufficient temporal context to give accurate forecasts. Figure 3 shows the analysis of the residual distribution. It demonstrates that the proposed model is robust with respect statistics. Our model is able to obtain a stable forecast of the water quality parameters which was made possible by centring of the error frequency around the zero-axis, thereby mitigating the systematic bias.

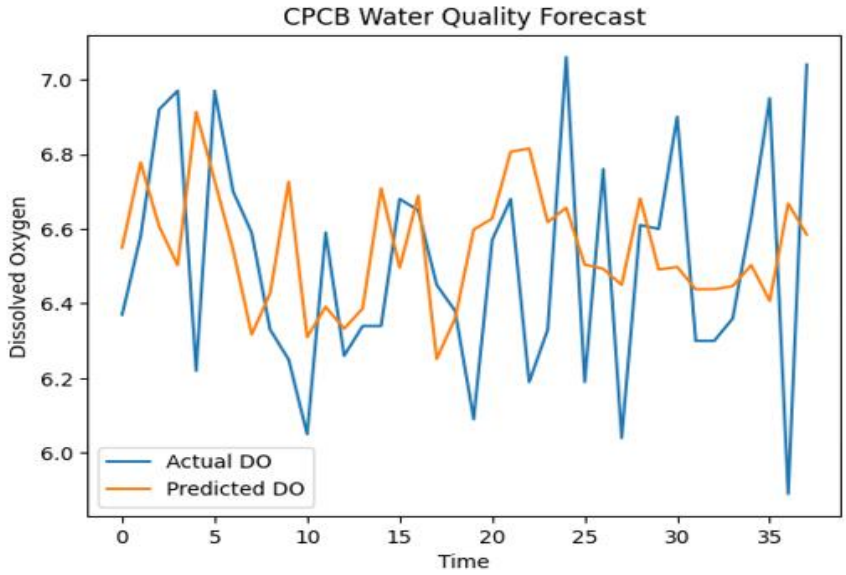

Figure 2: Comparison of Observed and Predicted Dissolved Oxygen (DO) Levels

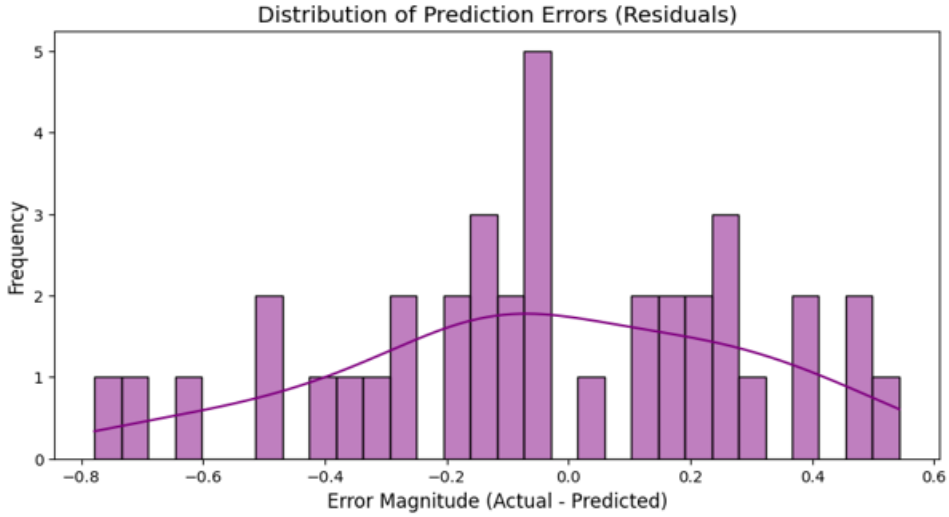

Figure 3: Distribution of Prediction Errors (Residuals)

Volume 65 | Issue 04 | April 2026
DOI: 10.5281/zenodo.19593147

Figure 4 shows the boxplot analysis which confirms the necessity for feature scaling. It highlights the huge differences in numerical scales between parameters like Conductivity and pH. In addition, the presence of distinct outliers demanded a robust modeling approach. This lead to the choice of the Attention-weighted hybrid framework.

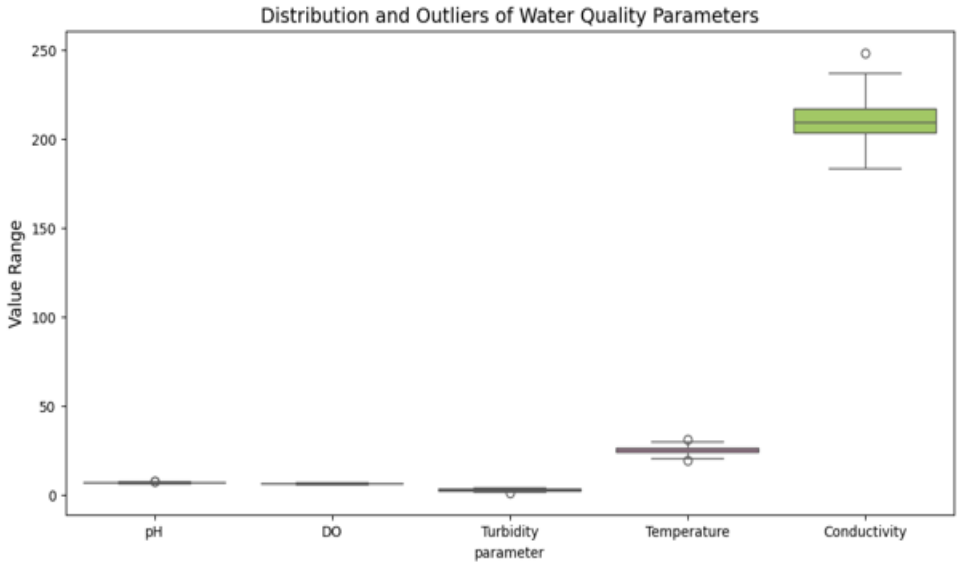

Figure 4: Statistical Distribution and Outlier Analysis of Water Quality Parameters

5. CONCLUSION

In this research, a hybrid BiLSTM-Attention-XGBoost architecture for forecasting urban water health with high precision was developed and validated. The proposed framework has merged the temporal memory of Bi-directional LSTMs with the robust regressive power of XGBoost, thereby it effectively addresses the gaps that prevails in standalone deep learning models when handling noisy environmental data. The proposed framework has achieved an RMSE of 0.3319 which assures that the system can reliably predict fluctuations in DO, and thereby serves as a main indicator of aquatic life sustainability. One of the key findings of the proposed work is the role of the Attention mechanism in prioritizing important shifts in parameters like pH and Conductivity, thereby enhancing the model's sensitivity to sudden pollution events. Our present work used historical CPCB records to demonstrate a proof-of-concept. The modular design of this proposed method allows for straightforward integration into live IoT-based sensor networks. Ultimately, this research provides a practical, repeatable tool for urban planners to anticipate degradation events and safeguard public water resources through an automated, data-driven early warning system.

References

1) Zhang, Y., Liu, X., & Chen, L. (2021). Deep learning-based water quality prediction using LSTM networks. *Journal of Hydrology*, *603*, 127053. doi.org

2) Wang, H., Zhao, J., & Li, Q. (2022). A hybrid LSTM–XGBoost model for water quality prediction. *Water Resources Management*, *36*(5), 1689–1703. doi.org

3) Li, F., Sun, W., & Zhang, H. (2024). Intelligent water quality prediction using a hybrid CNN–LSTM model. *Environmental Science and Pollution Research*, *31*, 11234–11248.

Volume 65 | Issue 04 | April 2026
DOI: 10.5281/zenodo.19593147

4) Chen, G., Xu, P., & Yang, D. (2024). Multi-source data fusion for water quality prediction using machine learning techniques. *Science of the Total Environment*, *912*, 168921.

5) Singh, R., & Kumar, M. (2023). Time-series forecasting of water quality parameters using machine learning approaches. *Journal of Environmental Management*, *330*, 117125.

6) Rahman, A., Patel, S., & Verma, R. (2026). Hybrid AI models for interpretable water quality prediction in data-scarce environments. *Environmental Modelling & Software*, *185*, 106012.

7) Wu, J., & Wang, Z. (2022). A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory. *Water*, *14*(4), 610. doi.org

8) Utku, A., Akpinar, E., & Gök, M. S. (2023). A CNN-LSTM hybrid model for water quality classification and assessment using sensor-based environmental data. *Environmental Monitoring and Assessment*, *195*(1), 122.

9) Luo, X., Zhang, H., & Liu, Y. (2024). Multi-step water quality forecasting using an encoder-decoder CNN-LSTM model with attention mechanism. *Journal of Hydrology*, *628*, 130456.

10) Wang, L., Chen, Y., & Zhao, X. (2024). Hybrid LSTM-GRU model with Bayesian optimization for watershed water quality prediction. *Environmental Science and Pollution Research*, *31*(12), 15890–15904.

11) Yan, T., Zhao, J., & Sun, L. (2024). A comprehensive review of machine learning-based water quality prediction techniques: Trends, challenges, and future directions. *Water Resources Management*, *38*(5), 1821–1845.

12) Bagheri, M., Mohammadi, M., & Sharifi, S. (2024). Hybrid CNN-LSTM with XGBoost feature importance for real-time dissolved oxygen prediction in urban water systems. *Process Safety and Environmental Protection*, *182*, 412–425.

13) Guo, H., Zheng, J., & Tan, Y. (2024). Intelligent water quality prediction system with a hybrid CNN–LSTM model. *Water Practice & Technology*, *19*(11), 4538–4552. doi.org

14) Dharmarathne, G., Punniedat, S., & Herath, H. (2025). A review of machine learning and internet-of-things on the assessment and monitoring of water quality. *Results in Engineering*, *26*, 105182. doi.org