IoT based real time Water Quality Forecasting using Machine Learning and Deep Learning Techniques

Dr. M. Geetha Jenifel[1], Dr. Mashael Mahmoud Khayyat[2]
[1]Postdoctoral Researcher, Centre of Postgraduate Studies, Lincoln University College, Malaysia;
[2]Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah;
pdf.jenifel@lincoln.edu.my, mkhayyat@uj.edu.sa

Abstract: Water quality forecasting plays a vital role in environmental sustainability, public health, and smart water resource management. The statistical methods commonly used are not adequate to encompass the nonlinear and dynamic nature of the data. Recently, machine learning (ML) and deep learning (DL) techniques have shown big improvements in predicting water quality parameters like pH, dissolved oxygen (DO), turbidity, biochemical oxygen demand (BOD), and water quality index (WQI). In this article, some of the latest ML and DL methods for water quality forecasting are introduced for the final implementation of the trustworthy and privacy-aware water quality forecasting using federated and explainable AI. The research findings are the IoT sensors enable continuous monitoring in multiple locations and reduce delays. The LSTM outperform traditional machine learning methods in forecasting. This work ensures that authorities and industries can make faster and more informed decisions regarding water treatment, pollution control and resource management.

Keywords: Public health; statistical models; water resources; prediction accuracy; environment

Introduction

Water is a precious natural resource, providing support to human life, agriculture, industry and environmental sustainability. Due to its rapid urbanization, industrial discharge, agricultural run-off, population growth and climate change, the quality of water around the world has gotten significantly worse. The World Health Organization (WHO) estimates that each year millions of people become sick due to waterborne diseases from contaminated water sources. Accordingly, water quality monitoring and correct prediction of water quality have evolved to be an essential field of study for environmental scientists, policymakers and smart city planners. Water quality forecasting involves forecasting the future condition of water bodies based on historical and current environmental conditions. The pH, dissolved oxygen (DO), turbidity, total dissolved solids (TDS), biochemical oxygen demand (BOD), chemical oxygen demand (COD), nitrate concentration, phosphate concentration, and water quality index (WQI) are all important water quality parameters. Precise forecasts allow for the implementation of early warning, pollution management, sustainable resource use and water treatment planning.

The application of statistical, mathematical models like Multiple Linear Regression (MLR), hydrological models or Autoregressive Integrated Moving Average (ARIMA) was the traditional approach to water quality prediction. These approaches work well in exploratory environments with small and linear datasets but are unable to recognize the nonlinear interdependence and temporal structure within environmental systems. The last few years have seen significant progress in Artificial Intelligence (AI), Deep Learning (DL) and Machine Learning (ML) technologies tackle the challenge of environmental monitoring and water quality forecasting. Through learning the hidden relationship from a large data set, ML algorithms can achieve more accurate prediction results than traditional algorithms, such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF) and Extreme Gradient Boosting (XGBoost). These algorithms can process nonlinear patterns, missing data, and environmental data with multiple dimensions. Hence, in this study, we shall review and analyze the latest Machine Learning and Deep Learning methods for water quality forecasting.

Related work

The growing demand for sustainable management of water resources, prevention of pollution and protection of public health has motivated the development of water quality forecasting as an active research field. The recent progress of Machine Learning (ML) and Deep Learning (DL) techniques has led to great improvements in forecasting accuracy with the ability to process nonlinear, high-dimensional and temporal environmental datasets. Various research groups have investigated various methods using AI to predict water quality parameters like pH, dissolved oxygen (DO), turbidity, biochemical oxygen demand (BOD), chemical oxygen demand (COD) and the Water Quality Index (WQI).

Table 1. Water quality comparative review

Ref	Techniques used	Dataset/Region	Key findings	Limitations
[1]	ML/DL (KNN Imputation, Noise Filtering)	Galicia Region, Spain	KNN imputation and noise filtering significantly enhanced prediction accuracy; smaller learning rates and optimized batch sizes improved model generalization on sparse environmental data.	Limited to a specific geographic region; generalizability to other environments not established.
[2]	Multi-View LSTM (MV-LSTM, MV-Online-LSTM)	Multiple monitoring stations (time series)	MV-Online-LSTM outperformed standalone LSTM models in accuracy and temporal dependency learning for multi-station water quality time series prediction.	Computational overhead of multi-view architecture not fully assessed; scalability to large-scale deployments remains unclear.
[3]	Statistical & ML: RF, SVM, XGBoost (Review)	Review-based (multiple datasets)	RF, SVM, and XGBoost were identified as the most accurate models for WQI prediction and pollution classification tasks.	Being a review study, no new experimental validation was conducted; dataset heterogeneity limits direct comparisons.

[4]	ML & Bayesian Deep Learning (uncertainty quantification)	Water resource management datasets	Bayesian deep learning models with uncertainty quantification proved more useful in critical water resource management applications.	Increased computational complexity due to Bayesian inference; may not be suitable for real-time applications without optimization.
[5]	LTSF-Linear (Lightweight Deep Learning)	Huangyang Reservoir, China	Lightweight LTSF-Linear model effectively forecast water quality parameters, demonstrating computational feasibility for real-time environmental monitoring.	Lightweight architecture may sacrifice prediction depth for complex, multi-parameter scenarios.
[6][7][8]	RF, SVM, and two additional advanced ML algorithms	Pollutant-focused datasets	RF and SVM demonstrated strong accuracy for predicting pollutants, particularly for nonlinear relationships between water quality parameters.	The study did not explore deep learning alternatives; performance on highly dynamic or noisy datasets was not assessed.
[9][10]	Deep Learning (data preprocessing, augmentation, prediction pipeline)	Multiple water quality datasets	A unified DL framework combining preprocessing, data enhancement, and prediction improved robustness and forecast stability across diverse datasets.	Framework complexity may limit interpretability; computational requirements for data enhancement not fully discussed.

Key Contribution

Recent studies have demonstrated that deep learning architectures like LSTM, GRU, CNN, and Transformer models can effectively learn temporal and spatial relationships in data related to the environment[11][12][13]. However, these developments come with a number of challenges yet to be addressed, such as the lack of data, model interpretability, uncertainty quantification, computational complexity, and privacy issues in distributed environmental monitoring systems[14][15].

The novelty of this work lies in

- The integration of IoT sensor networks with LSTM-driven forecasting techniques to develop a real-time smart river monitoring system capable of accurately predicting water quality trends and detecting pollution events before they become critical.
- The framework is specifically designed to support sustainable water resource management and proactive environmental monitoring in the Chittar Pattanam Channel region, enabling timely decision-making and improved ecosystem protection.

Method, Experiments and Results

Water quality forecasting predicts future values of important parameters such as pH, dissolved oxygen (DO), turbidity, biochemical oxygen demand (BOD), chemical oxygen demand (COD), temperature, nitrate concentration, and total dissolved solids (TDS). Figure 1 shows the model which helps governments, industries, and environmental agencies manage water resources and detect pollution early.

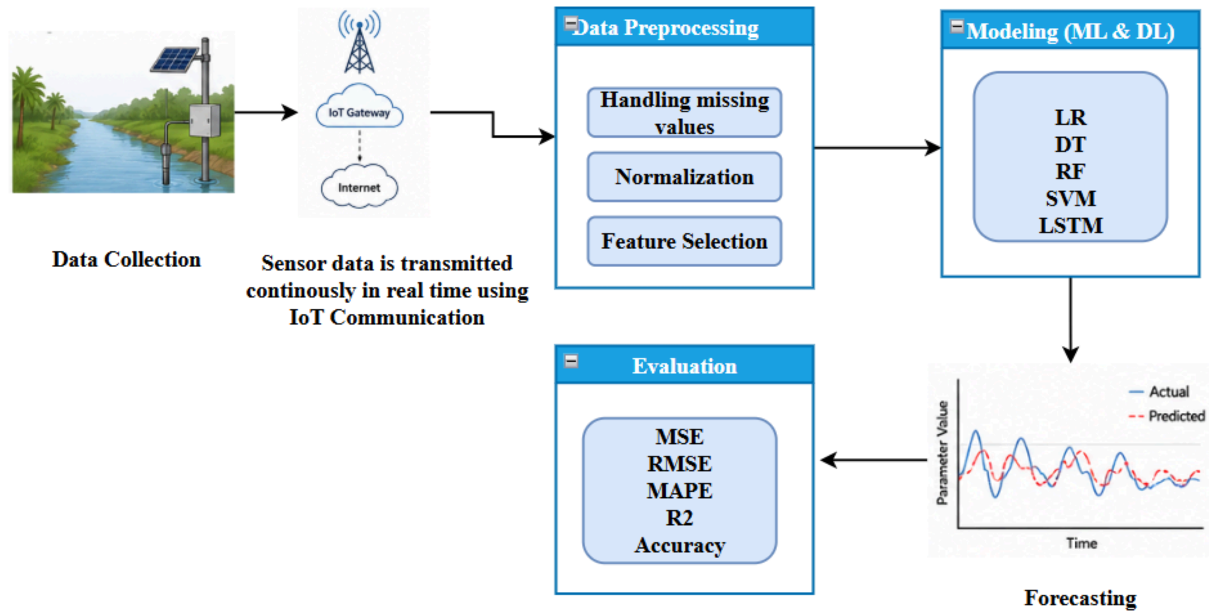

Figure 1. Real time water quality monitoring and forecasting

Machine Learning (ML) and Deep Learning (DL) techniques are highly effective because water quality data is: Nonlinear, Time-dependent, Multivariate and Noisy and seasonal.

Suppose water quality is measured over time. Let

$$X_t = [x_1, x_2, ...x_n] \qquad \text{----- (1)}$$

Where X_t=vector of water quality features at time t.x_i=parameters like pH, turbidity, DO and temperature.

The mail goal is to predict future water quality parameter. In the following equation f is the machine learning model and k is the historical lag window.

$$Y'_{t+1} = f(X_t, X_{t-1}, ...X_{t-k}) \qquad \text{----- (2)}$$

Data Collection

The water quality data used in this study is collected from the Chittar Pattanam Channel located in Kanyakumari District. This river channel plays a significant role in supporting agricultural activities, domestic water usage, and the surrounding ecosystem.

Missing Value Handling

Missing values are a common problem in real-time water quality monitoring systems because sensors may fail temporarily, communication networks may experience interruptions, or environmental disturbances can affect data transmission. Incomplete data can reduce the accuracy of Machine Learning and Deep Learning models; therefore, proper missing value handling is essential before model training.

$$x(t) = x_1 + \frac{(t-t_1)(x_2-x_1)}{t_2-t_1} \qquad \text{----- (3)}$$

Normalization

Normalization is a preprocessing technique used to scale water quality parameters into a uniform numerical range before applying Machine Learning and Deep Learning algorithms. Water quality attributes such as pH, turbidity, dissolved oxygen, and conductivity have different measurement units and value ranges. If the data are not normalized, parameters with larger numerical values may dominate the learning process, reducing model performance.

$$x^{'} = \frac{x-x_{min}}{x_{max}-x_{min}} \qquad \text{----- (4)}$$

Linear Models

Linear Regression is a supervised Machine Learning algorithm used to predict continuous numerical values by establishing a linear relationship between input variables and output variables. In water quality forecasting, it is used to estimate parameters such as pH, dissolved oxygen, or turbidity based on historical environmental data. The model attempts to fit the best straight line that minimizes the prediction error between actual and predicted values.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \epsilon \qquad \text{----- (5)}$$

In the above equation y is the predicted water quality parameter, β_i is the co efficients and ϵ is the error term.

Decision Tree Regression

Decision Tree Regression is a supervised Machine Learning technique used to predict continuous values by dividing the dataset into smaller subsets based on decision rules. The model forms a tree-like structure where each internal node represents a condition on an input feature and each leaf node represents a predicted output value. In water quality forecasting, Decision Tree Regression helps identify complex relationships between environmental parameters such as pH, turbidity, and dissolved oxygen.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - y_i^{'})^2 \qquad \text{----- (6)}$$

Random Forest

Random Forest is an ensemble Machine Learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Each decision tree is trained using randomly selected subsets of the dataset and input features, which increases model diversity and robustness. In water quality forecasting, Random Forest effectively captures nonlinear relationships among parameters such as pH, turbidity, dissolved oxygen, and temperature.

$$y^{i} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \qquad \text{----- (7)}$$

Support Vector Regression (SVR)

Support Vector Machine (SVM) is a supervised Machine Learning algorithm used for both classification and regression tasks. In water quality forecasting, SVM identifies the optimal decision boundary that minimizes prediction error while maximizing the margin between data points. The algorithm is highly effective for handling nonlinear environmental datasets through the use of kernel functions such as the Radial Basis Function (RBF).

$$y_i - (wx_i + b) \leq \epsilon$$

Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a specialized Recurrent Neural Network (RNN) designed to learn long-term dependencies in sequential data. It is highly suitable for water quality forecasting because environmental parameters change continuously over time and depend on previous observations. LSTM uses memory cells and gating mechanisms to selectively retain or forget information during the learning process. This architecture helps overcome the vanishing gradient problem commonly found in traditional RNN models. Due to its ability to model temporal patterns and nonlinear relationships, LSTM provides highly accurate predictions for real-time water quality monitoring systems.

Discussions

Several forecasting models including Linear Regression, Decision Tree Regression, Random Forest, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) were implemented and compared. The experimental results showed that Deep Learning models performed better than traditional Machine Learning approaches for time-series water quality prediction.

Table 2. Water quality model comparative using machine learning and deep learning

Model	MAE	RMSE	R^2 Score	Accuracy (%)
Linear Regression	0.91	1.24	0.78	78
Decision Tree	0.73	1.02	0.84	84
Random Forest	0.45	0.68	0.92	92
Support vector machine	0.52	0.74	0.90	90
LSTM	0.32	0.49	0.96	96

In figure 2, among all models, the LSTM model achieved the highest forecasting accuracy because it effectively captured temporal dependencies and nonlinear relationships present in real-time sensor data.

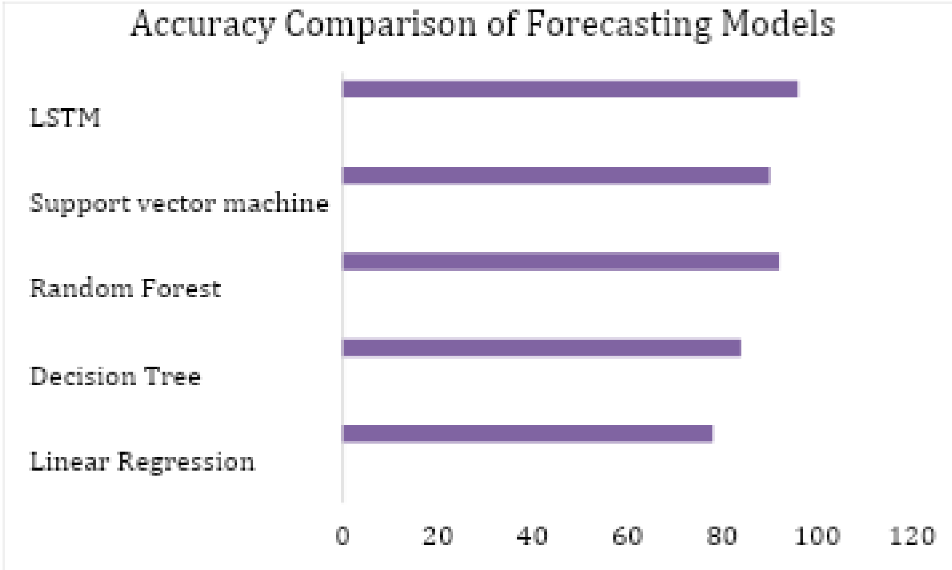

Figure 2. Accuracy comparison of machine learning and deep learning forecasting

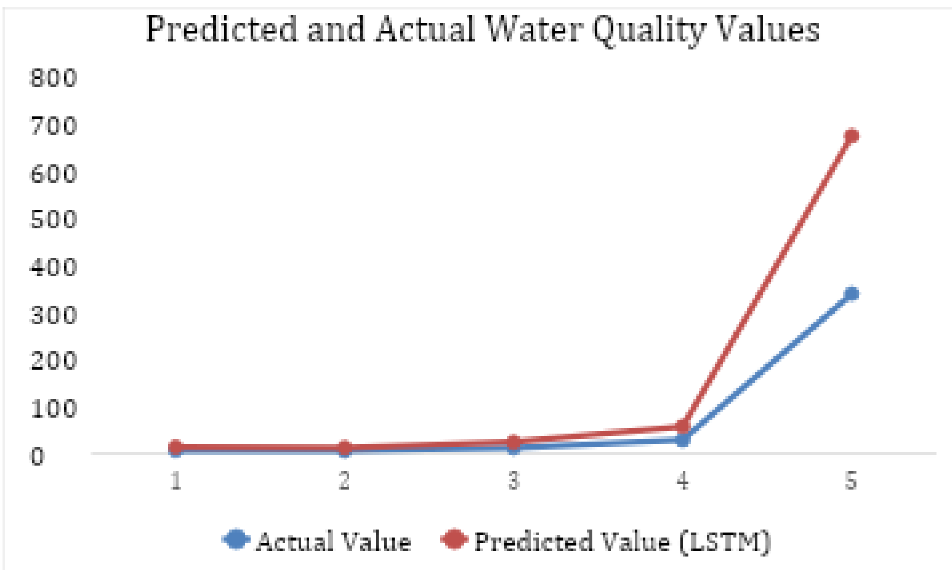

Figure 3. Predicted and actual values using machine learning and deep learning forecasting

The obtained results from the figure 3 confirm that Deep Learning models are more suitable for real-time water quality forecasting systems when compared with conventional Machine Learning techniques.

Conclusions

The results confirm that Deep Learning techniques,

- Particularly LSTM, provide highly accurate forecasting performance for real-time water quality monitoring applications.
- The integration of IoT sensor networks with intelligent forecasting models enables continuous environmental monitoring and supports early pollution detection.
- The developed forecasting framework can be effectively used for smart river monitoring systems and sustainable water resource management in the Chittar Pattanam Channel region.

This work is the preliminary phase of the trustworthy and privacy-aware water quality forecasting using federated and explainable AI. In future, privacy preserving technique will be incorporated.

References

1. A. Helaly et al., "Machine Learning and Deep Learning Techniques for Water Quality Prediction: A Practical Overview Using Galicia Region Data," Environmental Data Science, 2025.
2. X. Wang et al., "Multi-View Long Short-Term Memory (MV-LSTM) for Water Quality Prediction at Multiple Monitoring Stations," Water Resources Research, 2025.
3. N. Lokman et al., "A Detailed Review of Water Quality Forecasting and Classification Methods Employing Statistical and Machine Learning Techniques," Journal of Hydrology, 2025.
4. A. Mengistu et al., "Machine Learning and Bayesian Deep Learning Approaches for Water Quality Prediction and Uncertainty Quantification," Water Resources Management, 2026.
5. Y. Chen et al., "Long-Term Series Forecasting Linear (LTSF-Linear) Deep Learning Framework for Water Quality Parameter Prediction in Huangyang Reservoir," IEEE Access, 2024.
6. J. Han et al., "Predicting Water Quality Index Using Four Advanced Machine Learning Algorithms," Environmental Monitoring and Assessment, 2025.
7. L. Xu et al., "A Single Deep Learning-Based Water Quality Prediction System Integrating Data Preprocessing and Enhancement," Applied Soft Computing, 2024.
8. M. Huang et al., "Machine Learning for Water Quality Prediction in Sparse Datasets: Emphasis on Preprocessing, Interpolation, and Feature Engineering," Ecological Informatics, 2024.
9. M. Alwateer et al., "Explainable Machine Learning Models for Water Quality Prediction and Environmental Management," Science of the Total Environment, 2026.
10. S. Ismail et al., "Comparative Analysis of Water Quality Index Prediction Using Deep Learning Techniques: LSTM, TCN, ANN, and MLP," Neural Computing and Applications, 2024.
11. M. Zhu, J. Wang, and Y. Deng, “A review of the application of machine learning in water quality evaluation,” Environmental Research Communications, vol. 4, no. 11, pp. 1–24, 2022.
12. A. A. Hassan, M. A. M. Hassanien, and H. A. Hefny, “Water Quality Prediction Using Artificial Intelligence Algorithms,” Computational Intelligence and Neuroscience, vol. 2021, Art. no. 6659314, pp. 1–14, 2021.
13. M. Y. Shams, E. H. Houssein, M. A. Mahdy, and M. M. Elnahas, “Water quality prediction using machine learning models based on grid search method,” Multimedia Tools and Applications, vol. 83, pp. 1–29, 2024.
14. N. Nasir, N. A. M. Isa, and Z. Hussain, “Water quality classification using machine learning approaches and an intelligent real-time monitoring framework,” MethodsX, vol. 9, 2022.
15. W. C. Leong, A. Bahadori, J. Zhang, and Z. Ahmad, “Prediction of Water Quality Index (WQI) using Support Vector Machine (SVM) and Least Square-Support Vector Machine (LS-SVM),” International Journal of River Basin Management, vol. 19, no. 2, pp. 149–156, 2021.