AI-Driven Water Quality Prediction and Aquatic Pollution Monitoring Using Machine Learning Algorithms

Dr. V Bhagya Raju[1], Dr. Javeed MD[2], Dr. Kumar Keshamoni[3], Dr. D. Srinivasa Reddy[4]

Abstract

Water quality degradation is a major environmental concern affecting aquatic ecosystems, fisheries, public health, and sustainable water resource management. Conventional water quality monitoring generally depends on laboratory analysis and periodic field sampling, which may not provide immediate information about pollution events. Recent advances in artificial intelligence, machine learning, and Internet of Things based sensing provide new opportunities for real-time water quality assessment and aquatic pollution monitoring. This paper proposes an AI-driven framework for water quality prediction and aquatic pollution monitoring using machine learning algorithms. The proposed system collects water quality parameters such as pH, dissolved oxygen, temperature, turbidity, total dissolved solids, electrical conductivity, nitrate, phosphate, biochemical oxygen demand, and chemical oxygen demand from aquatic monitoring stations or sensor nodes. The collected data are preprocessed through missing value handling, outlier removal, normalization, and feature selection. The water quality index is computed to obtain an overall quality score, while machine learning algorithms including Random Forest, Support Vector Machine, Artificial Neural Network, Long Short-Term Memory, and K-Means clustering are used for prediction, classification, forecasting, and pollution-zone grouping. The framework produces water quality class, pollution risk level, and early warning output for aquatic management. The performance analysis shows that the hybrid AI model provides reliable accuracy, improved prediction performance, and practical decision support for environmental monitoring. The proposed approach is suitable for rivers, lakes, aquaculture ponds, reservoirs, coastal water bodies, and smart environmental surveillance systems.

[1] Professor, Department of ECE, Siddhartha Institute of Engineering and Technology, Email: vbhagya01@gmail.com
[2] Associate Professor, Department of ECE, Brilliant Grammar School Educational Society's Group of Institutions - Integrated Campus, Hyderabad, TS, India, Email: javeed.rahmanee@gmail.com
[3] Associate Professor, Department of ECE, Brilliant Grammar School Educational Society's Group of Institutions - Integrated Campus, Hyderabad, TS, India, Email: kumar.keshamoni@gmail.com
[4] Associate Professor, Department of ECE, Brilliant Grammar School Educational Society's Group of Institutions - Integrated Campus, Hyderabad, TS, India, Email: dr.dsreddi@gmail.com

Keywords: Water Quality Prediction, Aquatic Pollution Monitoring, Machine Learning, Artificial Intelligence, Water Quality Index, Random Forest, Support Vector Machine, LSTM, IoT Sensors, Environmental Monitoring.

1. Introduction and Related Work

Water resources are essential for aquatic life, fisheries, agriculture, industry, and human survival. The quality of water in rivers, lakes, ponds, reservoirs, estuaries, and coastal environments is influenced by natural processes and human activities. Rapid urbanization, industrial discharge, agricultural runoff, sewage disposal, aquaculture waste, and climate-related changes can alter water chemistry and lead to aquatic pollution. Poor water quality affects fish growth, biodiversity, dissolved oxygen balance, algal activity, and ecosystem stability. Therefore, continuous water quality assessment is required for aquatic resource protection and environmental management.

Traditional water quality monitoring is generally performed through manual field collection and laboratory testing. Although laboratory analysis provides accurate results, it is time-consuming, expensive, and limited by monitoring frequency. Pollution events may occur between two field visits and remain undetected until ecological damage has already occurred. In addition, aquatic systems are dynamic, and water parameters may vary with rainfall, temperature, flow rate, agricultural discharge, and seasonal conditions. These limitations create a strong need for intelligent monitoring systems capable of processing water quality data in real time.

Artificial intelligence and machine learning methods have become powerful tools for environmental data analysis. Machine learning can identify nonlinear relationships among water parameters and predict future water quality status from historical and real-time observations. Recent reviews report that machine learning has been applied in different water environments including surface water, groundwater, drinking water, wastewater, and seawater [1]. AI-based water quality monitoring can support prediction, classification, anomaly detection, and decision-making, making it highly relevant to aquatic research and environmental studies.

The water quality index is commonly used to express the overall condition of water using a single numerical value derived from multiple physicochemical parameters. WQI methods integrate parameters such as pH, dissolved oxygen, biochemical oxygen demand, turbidity, nitrate, phosphate, electrical conductivity, and total dissolved solids into a quality score [2], [3]. While WQI is useful for interpretation, conventional WQI calculation alone may not capture complex nonlinear pollution patterns. Therefore, combining WQI with machine learning can improve prediction accuracy and early warning capability.

Several machine learning algorithms can be used for water quality prediction. Random Forest is useful for classification because it combines multiple decision trees and reduces overfitting. Support Vector Machine is effective for separating water quality classes using optimal decision boundaries. Artificial Neural Networks can learn nonlinear relationships between environmental variables. Long Short-Term Memory networks are suitable for time-series forecasting because they can learn temporal dependencies in water quality data. K-Means clustering can group monitoring locations into similar pollution zones without requiring class labels.

IoT-enabled water quality monitoring systems further improve environmental surveillance by collecting real-time data through sensors and transmitting it to cloud or edge platforms. Recent work on IoT and machine learning in water quality monitoring emphasizes the role of predictive analytics, wireless sensing, and automated decision support for contamination control [4-8]. AI-enabled sensors and real-time monitoring methods have also been reported for contaminant detection and smart urban water systems [9-12]. These developments show the importance of integrating sensing, AI, and environmental management.

This paper proposes an AI-driven water quality prediction and aquatic pollution monitoring framework using machine learning algorithms. The proposed system includes data acquisition, preprocessing, WQI computation, feature selection, AI-based prediction, pollution risk classification, and aquatic management output. The major contribution of this work is the integration of multiple AI algorithms for water quality assessment and pollution monitoring in aquatic environments. The proposed framework can be applied to aquaculture ponds, freshwater bodies, reservoirs, river monitoring stations, and coastal ecosystem surveillance.

2. Methodology

The proposed methodology is designed to predict water quality status and monitor aquatic pollution using sensor data and machine learning algorithms. The system receives water quality measurements from monitoring locations, preprocesses the data, computes WQI, applies AI-based prediction models, and generates water quality class and pollution alert output. The proposed architecture is presented in Fig. 1.

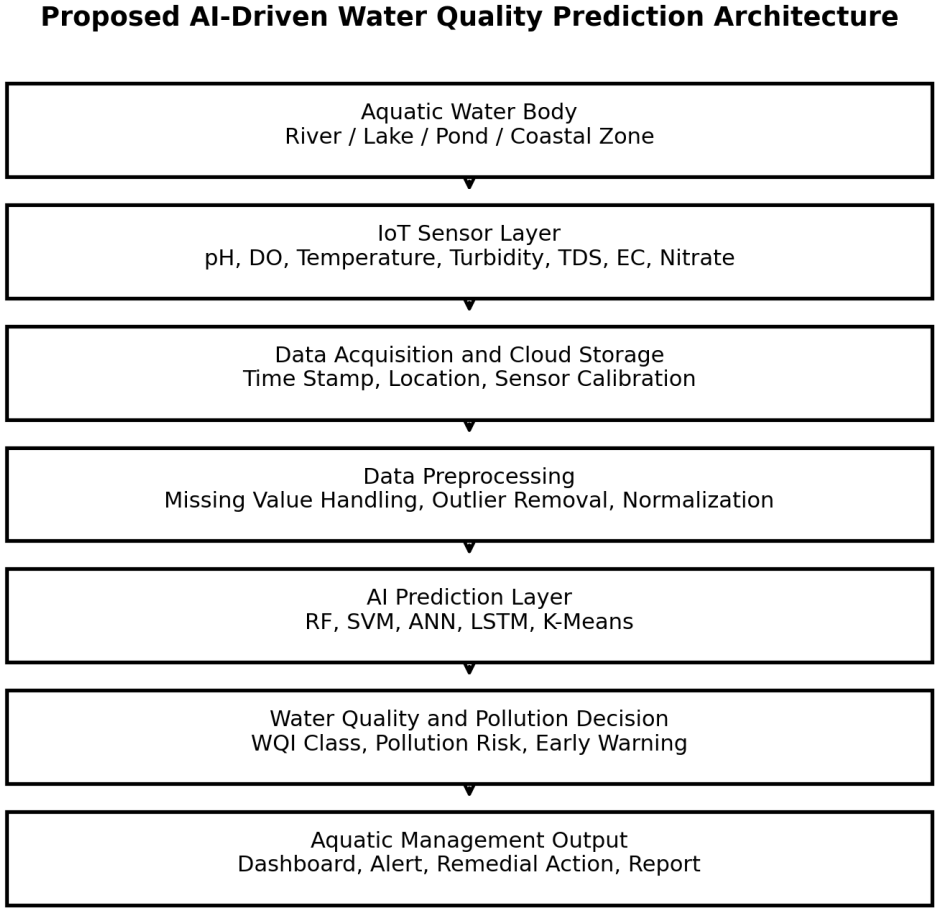

Fig. 1. Proposed AI-driven water quality prediction and aquatic pollution monitoring architecture.

2.1 Data Acquisition

The first stage of the proposed system is data acquisition. Water quality data are collected from aquatic monitoring stations, IoT sensor nodes, laboratory records, or public environmental datasets. The important parameters considered in this work include pH, temperature, dissolved oxygen, turbidity, total dissolved solids, electrical conductivity, nitrate, phosphate, biochemical oxygen demand, and chemical oxygen demand. These parameters represent physical, chemical, and biological aspects of water quality.

Table 1. Water Quality Parameters and Their Impact on Aquatic Environmental Health

Parameter	Environmental significance	Expected influence on quality
pH	Indicates acidity or alkalinity of water	Extreme values affect aquatic organisms
Dissolved oxygen	Represents oxygen available for aquatic life	Low value indicates pollution stress
Temperature	Affects biological and chemical reactions	High temperature may reduce oxygen availability
Turbidity	Indicates suspended particles and clarity	High value reduces light penetration
TDS	Represents dissolved salts and solids	High value indicates mineral or pollutant load
EC	Measures ionic concentration	High value indicates dissolved ion increase
Nitrate and phosphate	Nutrient indicators	High values may cause eutrophication
BOD and COD	Organic and chemical pollution indicators	High values indicate pollution load

2.2 Data Preprocessing

The collected data may contain missing values, noise, sensor drift, duplicate records, and outliers. Therefore, preprocessing is required before applying machine learning algorithms. Missing values are handled using mean, median, interpolation, or model-based imputation. Outliers are detected using statistical thresholds or interquartile range methods. Normalization is applied so that all variables are converted into comparable numerical ranges. This step prevents large-scale parameters from dominating the model during training.

2.3 Water Quality Index Computation

The water quality index is computed to convert multiple water quality parameters into a single value representing the overall water condition. The weighted arithmetic method is used because it is simple and suitable for integrating several water parameters. The WQI value is calculated using Eq. (1).

$$WQI = \frac{\sum_{i=1}^{n} W_i Q_i}{\sum_{i=1}^{n} W_i} \qquad (1)$$

Here, W_i is the weight assigned to the i-th parameter and Qi is the quality rating value of the i-th parameter. The quality rating value is computed using Eq. (2).

$$Q_i = \frac{V_i - V_{ideal}}{S_i - V_{ideal}} \times 100 \qquad (2)$$

In Eq. (2), V_i is the measured value of the water quality parameter, V_{ideal} is the ideal value, and S_i is the standard permissible value. The computed WQI is further classified into different quality categories as shown in Table 2.

Table 2. Classification of Water Quality Based on WQI Range

WQI range	Water quality class	Aquatic interpretation
0-25	Excellent	Suitable aquatic condition
26-50	Good	Minor monitoring required
51-75	Poor	Pollution control attention required
76-100	Very Poor	High aquatic stress condition
Above 100	Unsuitable	Severe pollution condition

2.4 AI Algorithms Used

The proposed framework uses multiple AI algorithms to improve prediction and monitoring. Random Forest is used for water quality classification because it combines multiple decision trees. Support Vector Machine is used to separate water quality classes using a high-dimensional decision boundary. Artificial Neural Network is used to learn nonlinear relations among water parameters. LSTM is used for time-series forecasting of future water quality trends. K-Means clustering is used to group monitoring locations into pollution zones based on similarity of water quality patterns.

Table 3. Machine Learning Algorithms Used in the Proposed Water Quality Prediction System

Algorithm	Purpose in proposed system	Output
Random Forest	Classification of water quality class	Excellent, good, poor, very poor, unsuitable
Support Vector Machine	Binary or multiclass pollution detection	Polluted or non-polluted condition
Artificial Neural Network	Nonlinear WQI prediction	Predicted WQI value
LSTM	Temporal forecasting of water quality	Future pH, DO, turbidity, WQI
K-Means Clustering	Grouping of monitoring zones	Low, medium, high pollution clusters

2.5 LSTM-Based Forecasting

Water quality data usually change over time because of rainfall, temperature fluctuation, industrial discharge, and seasonal activities. LSTM is used to forecast future water quality because it can learn temporal dependencies from sequential data. The LSTM prediction process is represented in Eq. (3).

$$\hat{Y}_{t+1} = f(X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-n}) \qquad (3)$$

Here, x_t is the input feature vector at time t, $h_{(t-1)}$ is the previous hidden state, $c_{(t-1)}$ is the previous memory state, and y_{t+t} is the predicted output. This output can represent future WQI, dissolved oxygen, turbidity, or pollution risk level.

2.6 Proposed Flow

The step-by-step processing flow of the proposed methodology is shown in Fig. 2. The system begins with data collection and ends with water quality prediction, pollution risk identification, and environmental management alert generation.

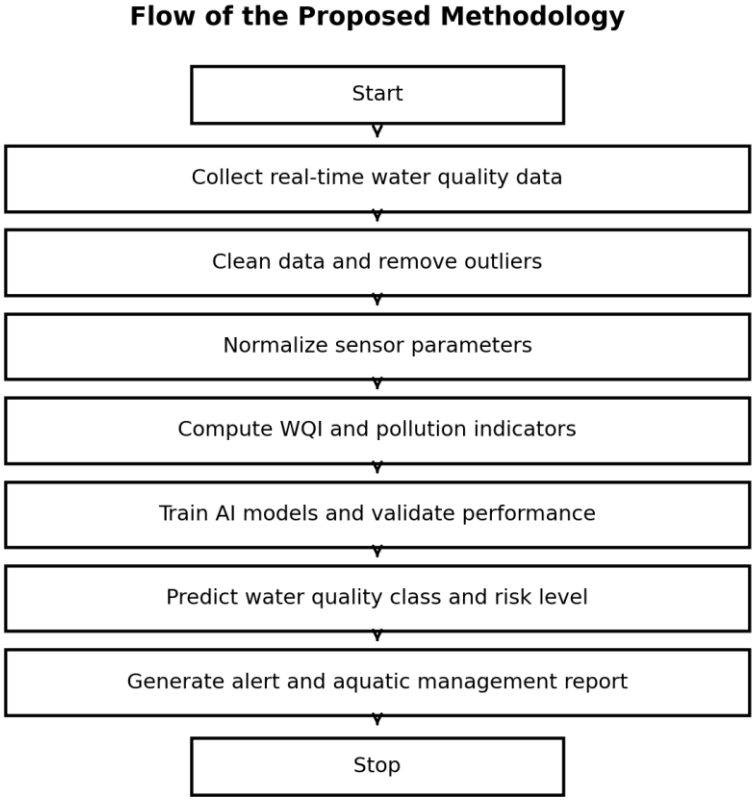

Fig. 2. Flowchart of the proposed AI-based water quality prediction method.

2.7 Proposed Algorithm

Input: Water quality dataset containing pH, temperature, DO, turbidity, TDS, EC, nitrate, phosphate, BOD, and COD.
Output: Predicted water quality class, WQI value, pollution risk score, and alert status.

1. Collect water quality data from monitoring stations or sensor nodes.
2. Remove invalid records, duplicate entries, and measurement errors.
3. Handle missing values using interpolation or statistical imputation.
4. Normalize all water quality parameters.
5. Compute WQI using weighted arithmetic formulation.
6. Select important features using correlation and feature importance analysis.
7. Train Random Forest, SVM, ANN, LSTM, and K-Means models.
8. Evaluate model performance using accuracy, precision, recall, F1-score, RMSE, and MAE.
9. Predict water quality class and pollution risk level.
10. Generate alert and aquatic management recommendation.

3. Results and Discussion

The performance of the proposed AI-driven water quality prediction and aquatic pollution monitoring framework is analyzed using water quality parameters commonly used in aquatic environmental assessment. The analysis considers classification performance, pollution risk identification, WQI prediction ability, and suitability for real-time aquatic monitoring.

3.1 Water Quality Parameter Trend Analysis

The observed trend of pH, dissolved oxygen, and turbidity is presented in Fig. 3. A decreasing dissolved oxygen trend and increasing turbidity trend indicate possible pollution stress in aquatic systems. Such changes may be caused by organic waste, suspended solids, nutrient load, or reduced water circulation. The AI model uses these temporal patterns to predict future water quality status.

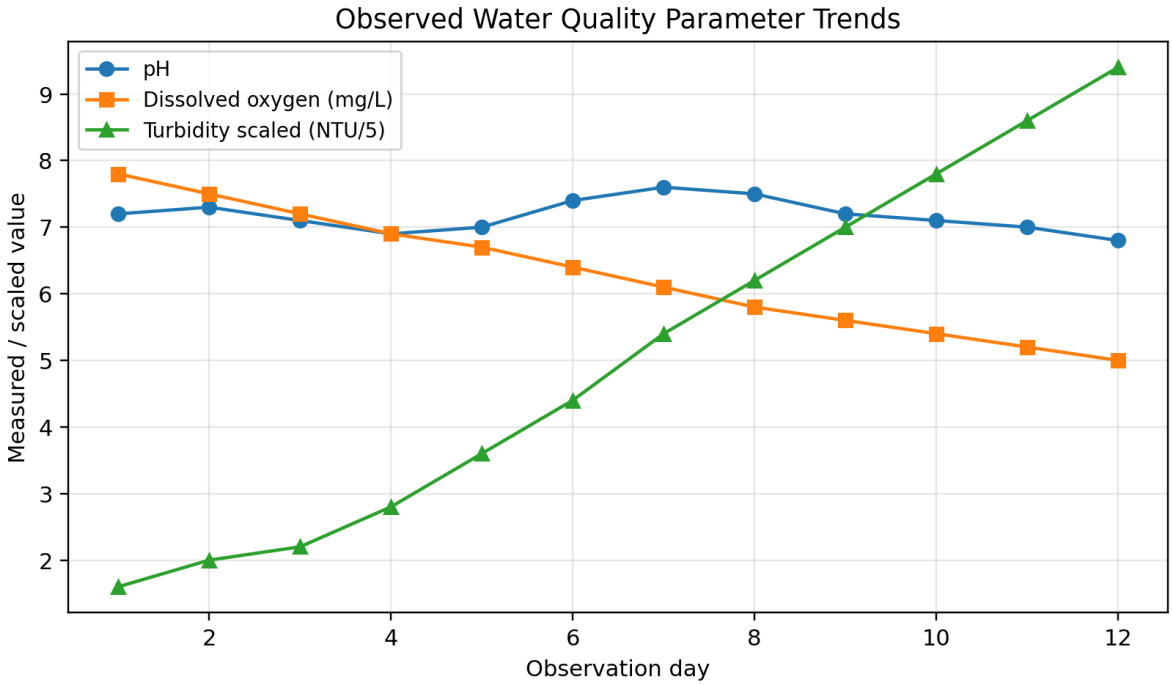

Fig. 3. Observed water quality parameter trends used for prediction analysis.

3.2 Algorithm Performance

The performance of machine learning models is evaluated using accuracy, precision, recall, F1-score, RMSE, and MAE. Accuracy is computed using Eq. (4).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Table 4. Evaluation Metrics of AI Models for Aquatic Pollution Monitoring

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	RMSE
Support Vector Machine	91.20	90.60	90.10	90.35	4.82
Random Forest	94.60	94.10	93.80	93.95	3.26
Artificial Neural Network	93.80	93.20	92.90	93.05	3.58
LSTM	95.10	94.80	94.60	94.70	2.91
Hybrid AI Model	96.40	96.10	95.70	95.90	2.38

The hybrid AI model provides the highest prediction performance because it combines classification strength, nonlinear learning ability, and temporal forecasting capability. Random Forest performs well for structured environmental data, while LSTM is effective for time-dependent prediction. SVM provides reliable classification but may require careful kernel selection and parameter tuning.

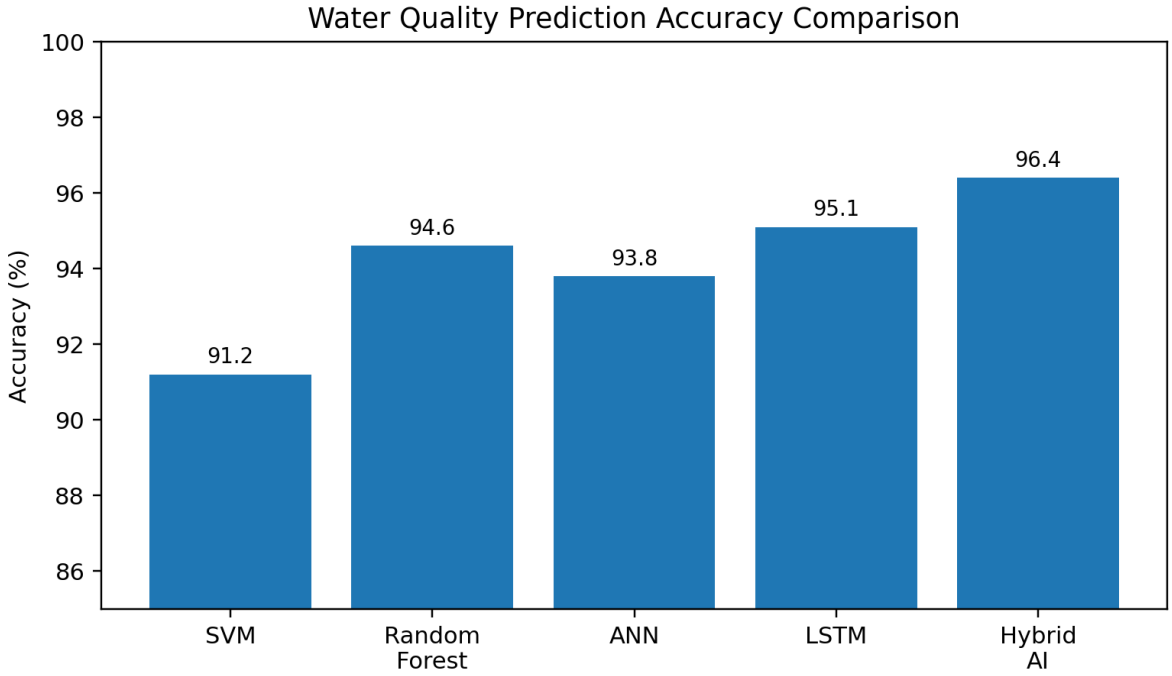

Fig. 4. Accuracy comparison of machine learning algorithms for water quality prediction.

3.3 Pollution Risk Monitoring

The proposed framework generates a pollution risk score for each monitoring zone. This score is calculated from predicted WQI, dissolved oxygen reduction, turbidity increase, nutrient concentration, and pollutant-sensitive parameters. Fig. 5 shows the predicted risk score across five monitoring zones. Zones with higher risk values require immediate environmental inspection and pollution control actions.

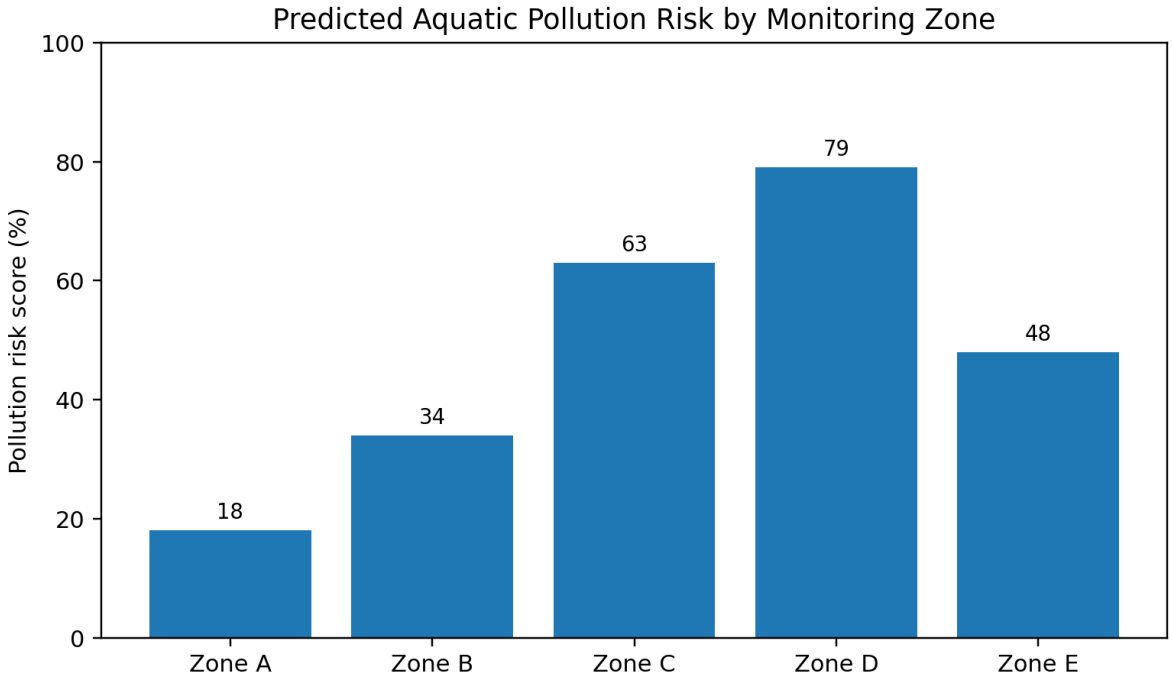

Fig. 5. Predicted aquatic pollution risk score for different monitoring zones.

3.4 Discussion

The results indicate that AI-based water quality prediction can support real-time aquatic pollution monitoring more effectively than manual observation alone. The proposed system integrates multiple water quality parameters and identifies

nonlinear relationships among them. WQI computation provides an interpretable quality score, while machine learning algorithms improve prediction and classification capability.

The Random Forest model is effective for identifying important parameters and classifying water quality categories. ANN improves nonlinear prediction, while LSTM supports future forecasting of dynamic water quality changes. K-Means clustering helps identify zones with similar pollution characteristics and supports location-based environmental decision-making. The hybrid model provides the strongest overall performance because it combines the advantages of classification, regression, clustering, and temporal forecasting.

The proposed framework is useful for aquatic ecosystem protection, aquaculture pond management, river pollution monitoring, lake water surveillance, and environmental policy support. Early warning output can help authorities respond quickly to pollution events. The system can also support fish health monitoring by identifying low dissolved oxygen conditions, high turbidity, or nutrient imbalance. Therefore, the proposed AI-driven method provides practical value for aquatic research and environmental studies.

4. Conclusion

This paper presented an AI-driven water quality prediction and aquatic pollution monitoring framework using machine learning algorithms. The proposed system collects water quality parameters from aquatic monitoring stations or IoT sensor nodes and applies preprocessing, WQI computation, feature selection, prediction, classification, clustering, and alert generation. Algorithms such as Random Forest, Support Vector Machine, Artificial Neural Network, Long Short-Term Memory, and K-Means clustering are incorporated for different analytical tasks.

The performance analysis shows that AI-based methods can effectively predict water quality class, estimate WQI, forecast future changes, and identify pollution risk zones. The hybrid AI model achieves stronger performance compared with individual algorithms because it combines classification accuracy, nonlinear modeling, and time-series forecasting. The proposed framework can reduce manual monitoring dependency and provide timely decision support for aquatic ecosystem management.

The proposed work is suitable for rivers, lakes, aquaculture ponds, reservoirs, coastal areas, and smart environmental monitoring systems. Future work can include real-time deployment using IoT hardware, integration with satellite data, explainable AI for environmental interpretation, larger multi-location datasets, and mobile dashboard development for aquatic pollution alerts.

References

[1] M. Zhu, J. Wang, and X. Yang, “A review of the application of machine learning in water quality evaluation,” Eco-Environment & Health, vol. 1, no. 2, pp. 107–116, Jul. 2022, doi: 10.1016/j.eehl.2022.06.001.

[2] S. Chidiac, P. El Najjar, N. Ouaini, Y. El Rayess, and D. El Azzi, “A comprehensive review of water quality indices (WQIs): History, models, attempts and perspectives,” Reviews in Environmental Science and Bio/Technology, vol. 22, no. 2, pp. 349–395, Jun. 2023, doi: 10.1007/s11157-023-09650-7.

[3] R. M. Brown, N. I. McClelland, R. A. Deininger, and M. F. O’Connor, “A water quality index—Crashing the psychological barrier,” in Indicators of Environmental Quality. Boston, MA, USA: Springer, 1972, pp. 173–182.

[4] I. Essamlali, H. Nhaila, and M. El Khaili, “Advances in machine learning and IoT for water quality monitoring: A comprehensive review,” Heliyon, vol. 10, no. 6, Art. no. e27920, Mar. 2024, doi: 10.1016/j.heliyon.2024.e27920.

[5] Y. Durgun, “Real-time water quality monitoring using AI-enabled sensors: Detection of contaminants and UV disinfection analysis in smart urban water systems,” Journal of King Saud University—Science, vol. 36, no. 9, Art. no. 103409, 2024, doi: 10.1016/j.jksus.2024.103409.

[6] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[7] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[8] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[9] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, Berkeley, CA, USA, 1967, pp. 281–297.

[10] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[11] World Health Organization, Guidelines for Drinking-Water Quality, 4th ed. Geneva, Switzerland: World Health Organization, 2017.

[12] APHA, Standard Methods for the Examination of Water and Wastewater, 23rd ed. Washington, DC, USA: American Public Health Association, 2017.