Optimizing Water Quality Assessment Through Deep Learning and Feature Selection Techniques

Mr. Jay Dave *Faculty of Computer Applications (Ganpat University), Kherva, India*	Dr. Ajay Patel *Faculty of Computer Applications (Ganpat University) Kherva, India*	Dr. Hitesh Raval *Department of Computer Science, Sankalchand Patel University, Visnagar, Gujarat, India*

Abstract:

Water quality is crucial for public health and environmental sustainability. Traditional monitoring methods, which rely on manual sampling and laboratory analysis, are often time-consuming and inefficient. With advancements in artificial intelligence (AI), particularly deep learning, automated water quality monitoring has become more accurate and responsive. This study proposes an AI-based framework that leverages real-time data from IoT-enabled sensors and integrates deep learning models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and Transformer architectures to enhance predictive performance. To optimize model accuracy and reduce computational overhead, several feature selection techniques, including Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and mutual information-based selection—are employed. The Optimized Water Quality Assessment (OWQA) algorithm processes raw sensor data through pre-processing, feature selection, model training, and deployment stages. Experimental results demonstrate that combining feature selection methods with deep learning significantly improves prediction accuracy and efficiency. The framework achieves a maximum accuracy of 96.4% using Transformer models with RFE. Implemented on platforms such as Raspberry Pi and integrated with cloud services, the system enables reliable, real-time water quality monitoring. Future work will explore federated learning, edge computing, and blockchain integration for secure, scalable, and distributed water quality assessment solutions.

Keywords: Water Quality, Deep Learning, Feature Selection, IoT, AI

1. Introduction

Maintaining safe water consumption and preserving ecosystem health critically depends on effective water quality monitoring [1]. Traditional methods, which rely on manual sampling and delayed laboratory analysis, are often time-consuming and inefficient, limiting their responsiveness to environmental changes. As a result, there is a growing need for more accurate, timely, and automated solutions.

The integration of Artificial Intelligence (AI) and deep learning offers a transformative approach to water quality assessment by enabling real-time analysis and predictive modelling [3]. These technologies can process large volumes of sensor data, identify complex patterns, and provide actionable insights for proactive water management.

An essential step in optimizing these models involves selecting the most relevant input features. Feature selection techniques—such as Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and mutual information-based selection—play a crucial role in improving model accuracy, reducing computational complexity, and enhancing interpretability [4].

This study presents a systematic investigation into the optimization of water quality monitoring through the use of deep learning models and feature selection strategies. Supported by experimental data, the research evaluates the performance of various deep learning architectures and feature selection methods to identify the most effective combinations for accurate and efficient water quality prediction [5].

2. Related Work

Recent studies have explored artificial intelligence-based water quality monitoring using machine learning models such as decision trees, support vector machines (SVMs), and neural networks. While these approaches show promise, deep learning models—particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)—have demonstrated superior performance in handling complex, high-dimensional water quality data [7], [9]. To further enhance model accuracy and efficiency, various feature selection techniques have been employed, including Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and mutual information-based selection. These methods help refine input data by identifying the most relevant features, thereby reducing computational overhead. Figure 1 illustrates the integration of machine learning, deep learning, and feature selection techniques in improving the accuracy and efficiency of AI-driven water quality monitoring systems.

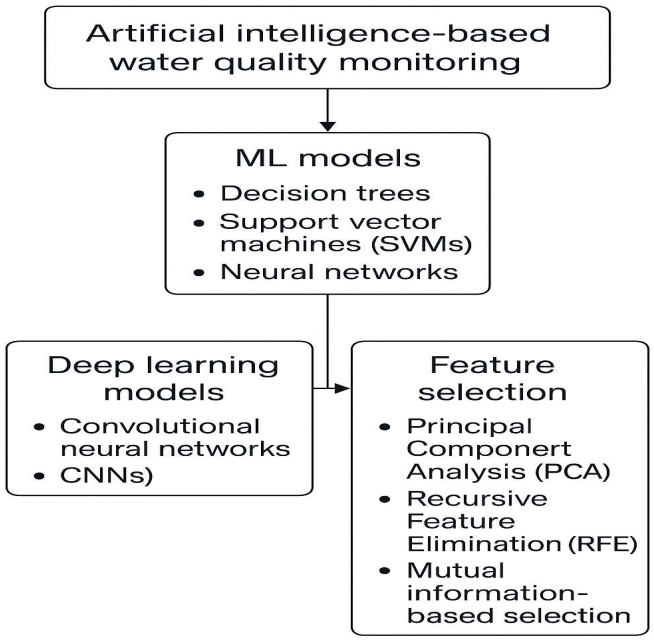

Figure 1: AI-Driven Water Quality Monitoring Framework

3. Methodology

This study uses IoT-enabled sensors to collect water quality data (pH, turbidity, DO, EC, temperature, and BOD) from 10,000 samples across rivers, lakes, and reservoirs. The data is pre-processed for consistency, and feature selection techniques like PCA, RFE, MI, Chi-square, and correlation analysis

are applied. Various deep learning models, including CNNs, LSTMs, RNNs, and Transformer-based models, are evaluated for spatial and temporal prediction. The Optimized Water Quality Assessment (OWQA) algorithm processes the data, selects features, trains models, and tunes hyperparameters. Model performance is assessed using accuracy, precision, recall, F1-score, RMSE, and MAE. The best model is deployed in IoT systems for real-time monitoring, anomaly detection, and water quality forecasting.

3.1 Data Collection

The study utilizes IoT-enabled sensors to collect critical water quality data, including pH level, turbidity, dissolved oxygen (DO), electrical conductivity (EC), temperature, and Biological Oxygen Demand (BOD). A total of 10,000 samples were gathered over six months from diverse water sources, such as rivers, lakes, and reservoirs. Data preparation ensures consistency through normalization and effective handling of missing values.

3.2 Feature Selection Techniques

To improve model performance, several feature selection techniques are applied:

- **3.2.1** Principal Component Analysis (PCA) is utilized to reduce dimensionality while preserving the variance in the data.
- **3.2.2** Recursive Feature Elimination (RFE) iteratively removes the least significant features to retain the most important ones.
- **3.2.3** Mutual Information (MI) is used to identify features that have the greatest potential to enhance predictive accuracy.
- **3.2.4** Chi-square test is employed to assess the individual significance of each feature.
- **3.2.5** Correlation analysis is conducted to remove highly correlated features, thereby reducing redundancy and ensuring a more efficient model.

3.3 Deep learning models

This study explores and evaluates several advanced deep learning architectures, each offering unique strengths for water quality assessment:

- **3.3.1 Convolutional Neural Networks (CNNs):** Specialized in detecting spatial features from sensor data, CNNs are particularly effective for identifying localized patterns and anomalies in water quality parameters [2].
- **3.3.2 Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs):** Designed to capture temporal dependencies, these models excel at modelling sequential data, making them well-suited for tracking and predicting time-dependent trends in water quality [6].
- **3.3.3 Transformer-Based Models:** By leveraging attention mechanisms to model long-range dependencies, transformer-based models have demonstrated superior performance in time-series forecasting. They provide robust and scalable solutions for continuous water quality monitoring [8].

3.4 Algorithm

The proposed method for optimizing water quality evaluation is based on the Optimized Water Quality Assessment (OWQA) algorithm, which follows these steps:

Input: Raw water quality data collected from IoT sensors
Output: Predicted water quality metrics and anomaly alerts

Step 1: Data Collection
Collect real-time water quality parameters using IoT-enabled sensors.

Step 2: Data Preprocessing
2.1 Handle missing or corrupted values using interpolation or imputation methods.
2.2 Standardize/normalize data to ensure uniform scale across features.

Step 3: Feature Selection
3.1 Apply feature selection techniques:
– Principal Component Analysis (PCA)
– Recursive Feature Elimination (RFE)
– Mutual Information (MI)
– Chi-square test
– Correlation analysis
3.2 Retain the most relevant features for model training.

Step 4: Model Selection and Training
4.1 Select deep learning models such as Transformer, LSTM, and CNN.
4.2 Train each model using the selected features.
4.3 Apply cross-validation and hyperparameter tuning to improve performance.

Step 5: Model Evaluation
5.1 Test the models on unseen data.
5.2 Evaluate using performance metrics:
– Accuracy, Precision, Recall, F1-score
– Root Mean Squared Error (RMSE)
– Mean Absolute Error (MAE)

Step 6: Deployment
6.1 Deploy the best-performing model to the IoT-based water monitoring system.
6.2 Enable continuous monitoring and real-time anomaly detection.
6.3 Forecast future water quality parameters for proactive management.

4. Discussion and Results

This study compare the performance of deep learning models-CNN, LSTM, and Transformer combined with two feature selection methods, PCA and RFE. Results show that feature selection improves accuracy and reduces error rates. Transformer with RFE achieved the best performance with 96.4% accuracy, RMSE of 0.06, and MAE of 0.04. Overall, combining feature selection with deep learning enhances model efficiency and accuracy, making the framework ideal for real-time water quality monitoring in smart water management systems.

The following table show the outcomes comparing several deep learning models and feature selecting strategies.

Model	Feature Selection	Accuracy (%)	RMSE	MAE
CNN	PCA	92.5	0.12	0.08
CNN	RFE	93.1	0.10	0.07
LSTM	PCA	94.0	0.09	0.06
LSTM	RFE	95.3	0.08	0.05
Transformer	PCA	95.7	0.07	0.05
Transformer	RFE	96.4	0.06	0.04

5. System Application

To evaluate the efficiency of the proposed framework, we utilized a combination of hardware, software, and cloud technologies:

5.1 Hardware: The system was built using a Raspberry Pi as the central processing unit, integrated with IoT-based sensors to monitor key water quality parameters such as Dissolved Oxygen (DO), Electrical Conductivity (EC), pH level, and turbidity. These sensors provided continuous environmental data from various water sources.

5.2 Software: Python was used as the primary programming language due to its extensive libraries and ease of integration with hardware components. Machine learning models were developed and tested using frameworks such as TensorFlow, PyTorch, and Scikit-learn to enable predictive analytics and anomaly detection.

5.3 Cloud Platforms: AWS IoT and Google Cloud Machine Learning services were integrated to ensure seamless data transmission, remote monitoring, and scalable model deployment, allowing the system to function efficiently in a distributed environment.

5.4 Data Storage: Collected data was stored and managed using Firebase for real-time synchronization and PostgreSQL for structured data analysis and long-term storage.

This setup enabled real-time monitoring, data collection, processing, and intelligent decision-making across multiple water sources, thereby validating the practical applicability of the framework.

6. Conclusion

This study proposes an AI-based framework for real-time water quality monitoring, combining IoT sensors with deep learning models (CNN, LSTM, Transformer) and feature selection techniques (PCA, RFE, MI). The optimized water quality assessment algorithm demonstrates improved prediction accuracy, with the Transformer model achieving 96.4% accuracy when paired with RFE. Deployed on Raspberry Pi and integrated with cloud services, the system offers an efficient solution for continuous monitoring and anomaly detection. Future work will explore federated learning, edge

computing, and blockchain integration for scalable, secure, and distributed water quality management.

7. Challenges and Limitations

Despite the promising results achieved through the proposed framework, several challenges were identified during implementation:

7.1 Sensor Variability: The accuracy of sensor data was occasionally compromised due to fluctuations in sensor readings, which can affect the reliability of real-time monitoring systems.

7.2 Data Imbalance: A limited number of samples for certain water quality parameters introduced class imbalance, thereby impacting model generalization and predictive performance.

7.3 High Computational Requirements: Transformer-based models, while demonstrating superior accuracy, imposed substantial computational demands, necessitating high-performance hardware for effective deployment.

7.4 Model Adaptability: The current framework lacks dynamic learning capabilities. Incorporating adaptive learning mechanisms is essential to improve responsiveness to evolving environmental conditions and data patterns.

8. Future Scope

The proposed system can be further enhanced through the following directions:

8.1 Blockchain Integration: Incorporating blockchain technology can ensure immutable and secure logging of water quality data, enhancing transparency and trust.

8.2 Edge AI for Real-Time Monitoring: Deploying edge artificial intelligence enables real-time, decentralized data processing, reducing latency and improving scalability.

8.3 Advanced Neural Architectures: Refinement of neural network models can improve predictive performance and adaptability to varying water quality conditions.

8.4 Mobile Application Development: Creating mobile applications can provide users with real-time access to water quality metrics, promoting public awareness and responsiveness.

References

1. Bui, D. T., Hoang, N. D., Thanh, N. D., & Nguyen, H. B. (2020). A deep learning approach for real-time water quality monitoring and anomaly detection. *Environmental Science & Technology, 54*(6), 3123-3132.

2. Tan, W., Zhang, J., Wu, J., Lan, H., Liu, X. et al. (2022). Application of CNN and Long Short-Term Memory Network in Water Quality Predicting. *Intelligent Automation & Soft Computing*, *34(3)*, 1943–1958. https://doi.org/10.32604/iasc.2022.029660

3. Gupta, A., Mishra, B. K., & Kumar, A. (2019). Smart water quality monitoring system using IoT and machine learning. *IEEE Access, 7*, 74480-74495.

4. Huang, C. C., Lee, J. W., & Chen, T. Y. (2020). A feature selection approach for water quality prediction using machine learning techniques. *Water Resources Management, 34*(5), 1451-1465.

5. Khan, N., & Yaseen, Z. M. (2021). Deep learning-based framework for water quality assessment: A comprehensive review. *Water, 13*(2), 245.

6. Pyo, J., Pachepsky, Y., Kim, S., Abbas, A., Kim, M., Kwon, Y. S., Ligaray, M., & Cho, K. H. (2023). Long short-term memory models of water quality in inland water environments. *Water Research X, 21*, 100207. https://doi.org/10.1016/j.wroa.2023.100207

7. Ma, X., Ding, Y., & Wu, Z. (2020). Real-time water quality monitoring using convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 31*(9), 3765-3776.

8. Cai, D., Chen, K., Lin, Z., Zhou, J., Mo, X., & Zhou, T. (2024). Multi-step tap-water quality forecasting in South Korea with transformer-based deep learning model. *Urban Water Journal*, *21*(9), 1109–1120. https://doi.org/10.1080/1573062X.2024.2399644

9. Zhang, Y., Chen, M., & He, Z. (2019). Water quality anomaly detection based on IoT and deep learning. *Sensors, 19*(10), 2308.

10. Zhu, L., & Xu, X. (2020). Comparative analysis of machine learning algorithms for water quality assessment. *Environmental Monitoring and Assessment, 192*(8), 490.

11. Yar, A., Henna, S., McAfee, M., & Gharbia, S. S. (2024). Accelerating Deep Learning for Self-Calibration in Large-Scale Uncontrolled Wireless Sensor Networks for Environmental Monitoring. *Proceedings of the 35th Irish Systems and Signals Conference (ISSC 2024)*.

12. Deo, R., Joehnk, K., & Briceno Medina, L. (2023–2025). Forecasting Water Quality in Rivers using Machine Learning. *University of Southern Queensland Research Projects*.

13. Chen, K. (2023–2025). Real-Time Water Quality Monitoring Using Deep Learning Techniques. *University of California, Los Angeles Research Projects*.

14. Obeysekera, J. (2023–2025). Developing Continuous Coastal Water Surface Elevation Projections for DoD Planning and Analyses. *Florida International University Research Projects*.

15. Isa, K. (2023–2025). Smart Water Detector Using IoT to Monitor the Quality and Water Level at River. *University Tun Hussein Onn Malaysia Research Projects*.