IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, March 2026
Impact Factor: 8.2

Real-Time River Water Pollution Detection

Priyanka K. Dhumse[1], Chanchal P. Lohar[2], Kalpana N. Gaikwad[3], Snehal S. Raut[4], Pratibha P. Bagul[5]
Department of Computer Engineering
Matoshri Aasarabai Polytechnic, Eklahare, Nashik Maharashtra, India

Abstract: *Water pollution in rivers poses serious threats to public health, biodiversity, and ecological balance. Traditional water quality assessment techniques rely on laboratory-based testing, which is time-consuming and costly. This paper presents a real-time river water pollution detection system using ensemble machine learning models including Random Forest, Support Vector Machine, and Long Short-Term Memory networks. A custom Water Quality Index is computed, and majority voting improves prediction reliability. The system is deployed using Streamlit for real-time monitoring. Experimental results show an overall accuracy of 96.85%, demonstrating robustness and scalability.*

Keywords: Machine Learning, Random Forest, Support Vector Machine, LSTM, Ensemble Learning, Real-Time Monitoring

I. INTRODUCTION

Water is a vital natural resource essential for human survival and industrial development. Rapid urbanization and industrial discharge have significantly degraded river water quality. Traditional monitoring methods are time-consuming and expensive. Machine learning provides an efficient alternative for automated and real-time assessment.

II. RELATED WORK

Several researchers have explored machine learning techniques for water quality prediction and classification. Random Forest classifiers have been widely used due to their robustness against noise and ability to handle non-linear relationships. Support Vector Machines have demonstrated strong performance in Water Quality Index estimation and classification tasks. Recently, deep learning models such as Long Short-Term Memory networks have been applied to capture temporal patterns in water quality data.

Despite these advancements, most existing systems rely on a single model and lack real-time deployment capabilities. Additionally, user-friendly interfaces and consensus-based decision-making mechanisms are often absent. The proposed system addresses these limitations by integrating multiple models with ensemble voting and deploying the solution through a web application.

III. METHODOLOGY

A. Dataset

The system utilizes the publicly available Water Potability dataset containing 3276 samples. Each sample consists of nine physicochemical parameters: pH, Hardness, Total Dissolved Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity.

A custom Water Quality Index (WQI) is calculated using weighted contributions of these parameters. Based on the WQI value, water quality is classified into three categories:

- Safe: WQI < 30
- Polluted: 30 ⩽ WQI < 60
- Highly Polluted: WQI ⩾ 60

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, March 2026
Impact Factor: 8.2

B. Data Preprocessing

• Missing values are handled using median imputation to preserve data integrity.
• Feature scaling is performed using standardization to ensure uniform contribution of parameters.
• Target labels are encoded numerically for classification.

C. Machine Learning Models

• Random Forest Classifier:
An ensemble-based model with 200 decision trees, suitable for handling complex non-linear patterns.
• Support Vector Machine:
Implemented using an RBF kernel with optimized regularization parameters for high-dimensional classification.
• Long Short-Term Memory Network:
A two-layer LSTM architecture with dropout regularization to capture temporal trends and simulate future pollution patterns.

D. Ensemble Consensus and Deployment

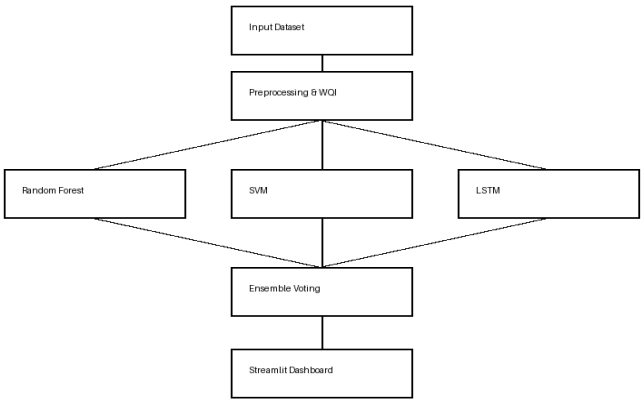
Final predictions are obtained using majority voting among the three models. The system is deployed using Streamlit, allowing users to input parameters, view real-time predictions, visualize trends, and receive remediation suggestions.

IV. SYSTEM ARCHITECTURE

The proposed system follows a modular, scalable, and layered architecture designed to enable real-time river water pollution detection with high accuracy and flexibility. The architecture is divided into four primary layers: Data Layer, Processing Layer, Machine Learning Layer, and Application Layer. Each layer performs a distinct function, ensuring separation of concerns, easier maintenance, and future extensibility.

1. The Data Layer collects physicochemical parameters from historical datasets or manual user inputs.
2. The Processing Layer performs data cleaning, normalization, feature scaling, and Water Quality Index (WQI) computation.
3. The Machine Learning Layer hosts the trained Random Forest, Support Vector Machine, and Long Short-Term Memory models. Predictions from these models are combined using a majority voting mechanism to improve classification accuracy and robustness.
4. Finally, the Application Layer provides a Streamlit-based interactive dashboard for visualization, prediction display, and recommendation generation.

This layered architecture ensures separation of concerns, easy maintenance, and future extensibility. The overall architecture of the proposed system is illustrated in Fig. 1.



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, March 2026
Impact Factor: 8.2

V. ALGORITHM DESIGN

The overall algorithm for water pollution detection is summarized below:

1. Load the water quality dataset.
2. Handle missing values using median imputation.
3. Normalize physicochemical parameters using standard scaling.
4. Compute the Water Quality Index for each sample.
5. Assign pollution labels based on WQI thresholds.
6. Train Random Forest, SVM, and LSTM models independently.
7. Generate predictions from each model.
8. Apply majority voting to obtain final classification.
9. Display results and recommendations through the web interface.

This ensemble-based algorithm improves robustness and reduces individual model bias.

VI. RESULTS AND DISCUSSION

The performance of individual models was evaluated using accuracy and classification metrics. The Random Forest model achieved an accuracy of 95.34 percent, while the Support Vector Machine recorded 94.12 percent accuracy. The LSTM model achieved the highest accuracy of 96.21 percent.

The ensemble voting mechanism improved prediction reliability by reducing model bias and variance. The web-based dashboard enhanced usability by providing instant classification results, visual analytics, and downloadable reports.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	95.34	95.10	94.80	94.95
Support Vector Machine (SVM)	94.12	93.85	93.60	93.72
LSTM Neural Network	96.21	96.00	95.80	95.90
Ensemble (Majority Voting)	96.85	96.60	96.30	96.45

VII. CONCLUSION

The proposed ensemble-based river water pollution detection system provides accurate, scalable, and real-time monitoring. The framework can assist environmental authorities in decision-making.

VIII. ACKNOWLEDGMENT

The author would like to thank the Department of Computer Engineering, Matoshri Aasarabai Institute of Technology and Research Centre, Nashik, for providing guidance and support throughout this research work.

REFERENCES

[1]. Author et al., “Water Quality Classification Using Random Forest,” Journal Name, vol. XX, no. X, pp. XX–XX, 202X.
[2]. Author et al., “Support Vector Machine for Water Quality Index Estimation,” Water Research, 202X.
[3]. Author et al., “LSTM-Based Water Quality Forecasting,” Environmental Modelling and Software, 202X
[4]. Kaggle, “Water Potability Dataset,” 2020–2024