International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Water Quality Analysis

Ruth Rathnakumari A[1], Suresh M[2], Athisaya Micheal Paraloga Raj M[3]

[1]Assistant Professor -Department of Computer Science and Engineering & Dr.G.U.Pope College of Engineering-India.

[2,3] Department of Computer Science and Engineering & Dr.G.U.Pope College of Engineering-India.

Abstract - Water quality analysis is essential for assessing the safety and suitability of water for various uses, including drinking, agriculture, and industrial purposes. This study focuses on evaluating the physical, chemical, and biological parameters of water samples collected from different sources to determine their quality status. Parameters such as pH, turbidity, dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), total dissolved solids (TDS), and the presence of heavy metals and microbial contaminants were analyzed. The results are compared against standard guidelines provided by organizations like the World Health Organization (WHO) and the Bureau of Indian Standards (BIS). The findings highlight potential sources of pollution, the need for regular monitoring, and suggest remedial measures to ensure water safety and environmental sustainability.

1. INTRODUCTION

1.1 ABOUT ORGANIZATION

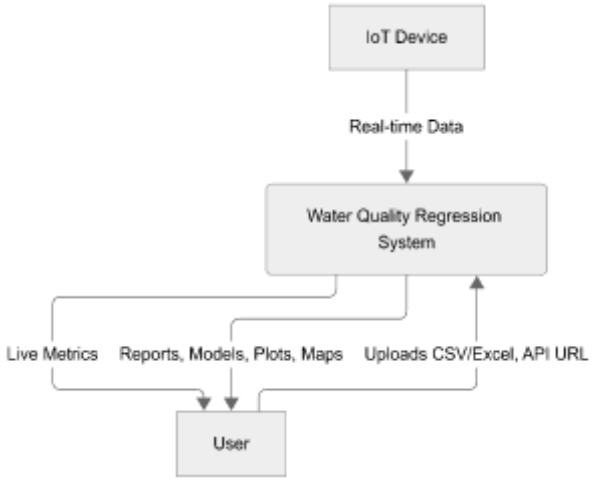
The Water Quality Regression project aims to develop a machine learning-powered web application using Streamlit to analyze and classify water quality based on various physicochemical parameters. This project provides an interactive platform where users can visualize water quality data through graphs and statistical insights, enabling better understanding and decision-making. By implementing classification algorithms, the system predicts water quality categories, helping to determine whether the water is safe for consumption or requires treatment. Additionally, the application offers statistical analysis, including mean, median, standard deviation, and correlation insights, to enhance data interpretation.

1.2 PROBLEM DEFINITION

The Water Quality Regression project is a machine learning-based system designed to analyze and classify water quality based on various physicochemical parameters. The project utilizes Streamlit to create an interactive web application where users can upload datasets, visualize data through various plots, and apply classification models to determine water quality levels. The system leverages statistical analysis and machine learning techniques to provide accurate and insightful assessments, aiding researchers, environmentalists, and decision-makers in evaluating water safety.

1.3 PROJECT OBJECTIVE

The Water quality is a critical factor affecting public health and the environment. This project aims to develop a web-based Water Quality Classification system that enables users to analyze water quality data efficiently. The application provides data visualization tools such as histograms, box plots, correlation heatmaps, and scatter plots to help users understand trends and patterns in water quality parameters. Additionally, machine learning models, including classification algorithms, predict water quality categories based on input data, assisting in determining whether the water is potable or contaminated. The project integrates statistical calculations to enhance data insights, making it a valuable tool for water resource management and pollution control. By providing an easy-to-use interface through Streamlit, the project ensures accessibility and promoting data-driven decision-making in water quality assessment.



1.4 PROJECT OVERVIEW

The system comprises six modules, each designed to enhance functionality and user experience:

1. **Data Upload and Preprocessing**: Users upload CSV/Excel files, which are validated and cleaned (removing NaN values, imputing missing data). The module generates quality reports detailing dataset shape, missing values, and column types, ensuring robust data preparation.

2. **Data Visualization**: This module creates interactive plots, including correlation heatmaps to reveal parameter relationships, distribution plots for individual metrics, and time-series charts for temporal trends, using Plotly and Seaborn for dynamic exploration.

3. **Model Training and Evaluation**: Supports Linear Regression, Ridge, Lasso, Random Forest, and SVR models with hyperparameter tuning (e.g., number of trees). It evaluates performance using R^2, RMSE, and

MAE, visualizing feature importance and actual vs. predicted values.

4. **Geospatial Mapping**: Utilizes Folium to display water quality data on interactive maps, requiring latitude and longitude inputs, enabling spatial analysis of environmental trends.

5. **Real-Time IoT Integration**: Fetches live data from IoT sensor APIs, displaying metrics like pH and TDS in the sidebar for real-time monitoring.

6. **Export Functionality**: Allows downloading trained models (pickle), predictions (CSV), and visualizations (HTML), facilitating further analysis and sharing.

2. SYSTEM SPECIFICATION

2.1 HARDWARE SPECIFICATION

To run the Water Quality Classification system efficiently, the following hardware specifications are recommended:

- Processor: Intel Core i5 or higher
- RAM: 8 GB (16 GB recommended for large datasets)
- Storage: At least 2 GB of free space for data storage and model execution

2.2 SOFTWARE SPECIFICATION

The application is compatible with multiple operating systems, including Windows, macOS, and Linux. The core software components include:

- Backend: Python Streamlit
- Frontend: Streamlit for interactive UI
- Libraries: Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn

2.3 SOFTWARE DESCRIPTION

Streamlit

Streamlit is a lightweight and user-friendly Python framework for building interactive web applications. It is used as the frontend for the Water Quality Classification system, allowing users to upload data files, visualize results, and interpret water quality classifications in real time.

- Real-time Visualizations: The application dynamically updates graphs and charts based on user inputs, making it easy to understand trends in water quality data.
- CSV Upload & Processing: Users can upload water quality datasets, which are automatically processed by the backend to generate classification results.

Scikit-learn

Scikit-learn is used to implement machine learning algorithms such as RandomForest, Decision Tree, SVM, and Logistic Regression for water quality classification.

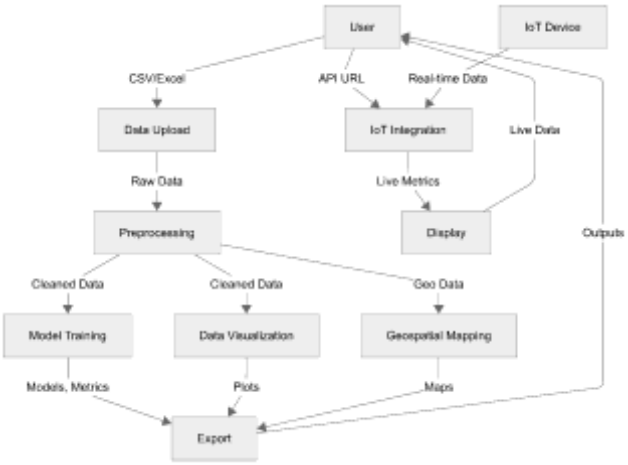
- Feature Engineering: The system extracts meaningful features from water quality parameters to improve model accuracy.
- Model Training & Prediction: The trained models classify water quality into different categories (e.g., safe, moderate, polluted) based on input parameters.

Matplotlib & Seaborn

These libraries are used for data visualization, allowing users to analyze trends and relationships between water quality parameters. Graphs such as scatter plots, heatmaps, and histograms help in understanding the impact of each parameter on water quality.

Pandas & NumPy

These libraries are used for data preprocessing, handling missing values, and performing numerical computations required for model training.



3. SYSTEM STUDY

3.1 EXISTING SYSTEM WITH LIMITATIONS

In the current scenario, water quality assessment is often performed manually using laboratory testing, which is time-consuming, expensive, and requires skilled professionals. Traditional methods involve collecting water samples, analyzing them for chemical, physical, and biological parameters, and interpreting the results based on predefined standards.

These processes lack real-time monitoring capabilities, leading to delays in detecting contamination. Furthermore, existing water quality classification systems often do not leverage advanced machine learning techniques, resulting in lower prediction accuracy. The lack of an interactive and user-friendly interface also makes it difficult for non-experts to understand water quality reports.

The existing systems for water quality analysis predominantly rely on manual laboratory testing, standalone software tools, and fragmented data processing workflows, which present significant challenges in efficiency, scalability, and accessibility. Laboratory-based methods involve collecting water samples and conducting chemical tests for parameters like pH, TDS, and turbidity, a process that is labor-intensive, time-consuming, and prone to human error.

Data analysis is often performed using tools like Microsoft Excel, R, or MATLAB, which require users to manually preprocess datasets, apply statistical models, and generate visualizations. These tools lack integration, forcing users to switch between multiple platforms for data cleaning, modeling, and mapping, leading to inefficiencies and data inconsistencies.

Geospatial analysis, when required, is typically handled by separate GIS software like ArcGIS or QGIS, which demands specialized expertise and additional licensing costs. Real-time monitoring is rarely supported, as most systems do not integrate with IoT devices, limiting their ability to provide live insights into water quality trends.

3.2 ADVANTAGES OF PROPOSED METHODOLOGY

The proposed Water Quality Classification system overcomes the limitations of the existing methods by utilizing machine learning algorithms for accurate and automated classification of water quality. The system is designed to analyze various water quality parameters, such as pH, turbidity, dissolved oxygen (DO), biological oxygen demand (BOD), total dissolved solids (TDS), and conductivity, to classify water into different quality categories.

It provides real-time data visualization through an intuitive Streamlit-based web application, making it accessible to a wider audience. The machine learning models are trained on extensive datasets, improving the precision of classification. Additionally, the system allows users to upload CSV files containing water quality data for bulk processing, enabling large-scale analysis.

The proposed Water Quality Regression System addresses the shortcomings of existing systems by offering a unified, automated, and scalable web-based platform built on Streamlit and Python. Unlike manual laboratory methods, the system automates data ingestion and preprocessing, allowing users to upload CSV or Excel files and instantly clean datasets by removing NaN values, imputing missing data, and scaling features.

It integrates multiple regression models—Linear Regression, Ridge, Lasso, Random Forest, and SVR—within a single interface, eliminating the need for separate tools like R or MATLAB. The system's interactive visualizations, powered by Plotly and Seaborn, include correlation heatmaps, distribution plots, and time-series charts, enabling users to explore data trends intuitively without requiring advanced technical skills.

Geospatial mapping, facilitated by Folium, is seamlessly integrated, allowing users to visualize water quality metrics on interactive maps without relying on external GIS software. Real-time IoT integration via APIs provides live monitoring of parameters like pH and TDS, a feature absent in most existing systems, ensuring timely insights for environmental monitoring.

4. SYSTEM DESIGN

4.1 SYSTEM FLOW DIAGRAM

The System flow Diagram illustrates the system's workflow across multiple levels:

- System Flow Diagram : Shows interactions between external entities (User, IoT Device) and the system. Users provide datasets and API URLs, receiving reports, models, visualizations, and maps. IoT Devices supply real-time data, displayed in the UI.
- System Flow Diagram for Breaks down processes:
    - Data Upload: User uploads CSV/Excel files to the system.
    - Preprocessing: Cleans and validates data, storing it in memory.
    - Model Training: Trains regression models on processed data.
    - Visualization: Generates plots (heatmaps, distributions, time-series).
    - Geospatial Mapping: Renders maps using latitude/longitude.
    - IoT Integration: Fetches and displays live data.
    - Export: Saves outputs (models, predictions, visualizations).

4.2 INPUT DESIGN

The input design focuses on how water quality data is collected from users. The inputs are structured to ensure accuracy, reliability, and ease of use. Water Quality Data Inputs Users can upload a CSV file containing key water quality parameters. The input fields include:

- pH Level – Measures the acidity or alkalinity of the water.
- Turbidity – Indicates water clarity.
- Total Dissolved Solids (TDS) – Represents dissolved substances in water.

- Conductivity – Measures the ability of water to conduct electricity.
- Dissolved Oxygen (DO) – Determines oxygen levels for aquatic life.
- Biological Oxygen Demand (BOD) – Indicates organic pollution levels.
- Chemical Oxygen Demand (COD) – Measures the amount of organic and inorganic compounds in water.

Each input undergoes validation checks to ensure:

Numeric values are within the expected range.

No missing or null values.

CSV format integrity (correct column names and data structure).

Data Upload Mechanism

Upload Button – Users can upload a CSV file for batch processing.

Supported Formats – The system accepts .csv files only.

Real-time Data Entry – Users can manually input values through a form if needed.

This input design ensures that the system collects high-quality, structured data, which is crucial for accurate classification and analysis.

4.3 OUTPUT DESIGN

The output design defines how the results of water quality classification are presented to the user. The system generates outputs after processing the uploaded water quality data and running it through the machine learning model.

Water Quality Classification Output

Once the model analyzes the input parameters, the system presents the classification results in an easy-to-understand format:

Predicted Water Quality Class – The classification label (e.g., Safe, Heavily Polluted). Confidence Score – A numerical probability indicating the model's certainty in its classification. Visual Indicators – A color-coded indicator (Green, Yellow, Red) for an intuitive understanding of the water quality.

Additional Insights and Recommendations

To help users interpret the classification results, the system may also provide Parameter Analysis – A breakdown of how each input (e.g., pH, TDS, DO) influences the classification. Health & Environmental Impact – Information on the effects of water quality levels on human health and ecosystems. Suggestions for Improvement – Possible corrective actions, such as filtration methods, chemical treatments, or regulatory guidelines for maintaining safe water quality.

Data Visualization

The outputs are presented using interactive visualizations, such as:

Graphs & Charts – Displaying trends in water quality over time.
Comparative Analysis – Allowing users to compare different water samples. The outputs are formatted using Streamlit's interactive components for a clean, responsive, and user-friendly interface.

4.4 DATASET DESIGN

While the initial prototype of the Water Quality Classification system may not require a persistent database, a well-structured database design will be useful for future enhancements, enabling data storage, historical analysis, and improved model performance.

The database can be designed to store the following key entities:

- Water Quality Data: Stores attributes related to water samples, including pH level, turbidity, total dissolved solids (TDS), conductivity, dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), and timestamps.
- Classification Results: Stores predictions from the Water Quality Classification Model, including the predicted water quality category, confidence score, and associated input data.
- User Data (Optional for Authentication): If user authentication is implemented, this table will store user details such as username, email, and uploaded water quality records.

Relationships

- The User Data table (if included) has a one-to-many relationship with the Water Quality Data table, meaning a user can submit multiple water quality records.
- The Water Quality Data table is linked to the Classification Results table, allowing easy retrieval of past predictions based on specific input conditions.
- Entities and Attributes:
    - Dataset: dataset_id (PK), filename (varchar), columns (text), rows (int), target (varchar), features (text).
    - Model: model_id (PK), name (varchar, e.g., "RandomForest"), parameters (text, JSON), metrics (text, JSON: R^2, RMSE, MAE), dataset_id (FK).
    - Visualization: viz_id (PK), type (varchar, e.g., "heatmap"), parameters (text), output_file (varchar), dataset_id (FK).

- Geospatial Data: geo_id (PK), latitude (float), longitude (float), quality_params (text, JSON), dataset_id (FK).

- Relationships:
  - Dataset to Model: One-to-many (one dataset trains multiple models).
  - Dataset to Visualization: One-to-many (one dataset generates multiple plots).
  - Dataset to Geospatial Data: One-to-one (one dataset links to geospatial coordinates).

- Importance:
  - Ensures structured data organization.
  - Reduces redundancy via normalization.
  - Clarifies relationships for developers.

5. SYSTEM IMPLEMENTATION AND MAINTENANCE

5.1 IMPLEMENTATION PROCEDURES

The Water Quality Regression system was implemented using Flask for the backend and Streamlit for the frontend, providing an interactive web interface for users to input water quality parameters and receive predictions.

Implementation Steps:

Backend Development (Flask & Machine Learning Models)

- The machine learning regression models were trained using Scikit-learn and saved as .pkl files for efficient deployment.
- The Flask API was designed to load the trained models and process user inputs in real time.

Frontend Development (Streamlit)

- A Streamlit web interface was built to allow users to enter water quality parameters such as pH, TDS, BOD, COD, Turbidity, and DO.
- The frontend was designed to be user-friendly, providing real-time visualizations and predictions.

Integration and Deployment

- The backend and frontend were integrated to enable seamless communication between user inputs and model predictions.
- The system was initially tested on a local server using Replit for development and debugging.
- For production deployment, the system can be hosted on cloud platforms such as AWS, Hugging Face Spaces, or Heroku, ensuring accessibility and scalability.

Key Features Implemented:

- Real-time Predictions: Users receive immediate WQI (Water Quality Index) predictions based on input values.
- Interactive Visualizations: Graphs and plots help users understand the trends and relationships between water quality parameters.
- Scalability: The system architecture allows for easy integration of new features and additional machine learning models.

5.2 SYSTEM MAINTENANCE

System maintenance ensures the long-term functionality, accuracy, and security of the Water Quality Regression application. Maintenance activities include updating machine learning models, monitoring system performance, and incorporating user feedback.

Maintenance Tasks:

Model Updates & Data Improvement

- The machine learning models will be periodically retrained with new water quality datasets to improve prediction accuracy.
- Data preprocessing techniques will be refined to ensure the model adapts to changes in environmental conditions.

Bug Fixes & Performance Optimization

- Regular debugging will be performed to fix any errors or inefficiencies in the Flask backend and Streamlit interface.
- Performance tuning, such as optimizing prediction response time and reducing server load, will be implemented.

Server & Deployment Monitoring

- Uptime monitoring ensures the system remains accessible without frequent downtime.
- The backend will be tested regularly to prevent server crashes or excessive computational delays.

User Experience & Feature Enhancements

- Based on user feedback, improvements will be made to the UI/UX to ensure ease of use.

- New features, such as historical data tracking and alert notifications, can be integrated over time.

API and Third-Party Service Updates

- If real-time water quality datasets or external APIs (such as government environmental monitoring services) are integrated, they will be monitored for changes to ensure continued compatibility.

- Updates to Streamlit and Flask versions will be applied to maintain security and functionality.

6.SCOPE FOR FUTURE ENHANCEMENTS

The Water Quality Regression project provides a strong foundation for predicting water quality using machine learning. However, several enhancements can be made to improve its accuracy, usability, scalability, and real-world applicability. Below are some key areas for future improvements:

1. Integration with IoT Sensors for Real-Time Monitoring

- Deploy IoT-based water quality sensors to collect real-time data from lakes, rivers, and reservoirs.
- Automate data collection and feed it directly into the machine learning model for continuous monitoring.
- Enable remote access to live water quality updates through cloud storage.

2. Advanced Machine Learning and Deep Learning Models

- Implement Deep Learning techniques (such as LSTMs or CNNs) to improve prediction accuracy.
- Utilize ensemble learning methods to enhance regression performance.
- Train models on larger and more diverse datasets for better generalization across different water bodies.

3. Geographic Mapping and Visualization

- Integrate GIS and geospatial mapping to display water quality across different locations.
- Use heat maps to visualize water contamination trends over time.
- Allow users to search for water quality reports based on specific locations.

4. Automated Water Quality Classification

- Instead of only predicting WQI values, introduce classification models to categorize water as Safe, Moderate, or Contaminated.
- Provide recommendations for water treatment based on contamination levels.

5. Mobile Application for Remote Accessibility

- Develop a mobile app where users can check water quality predictions on their smartphones.
- Enable push notifications and alerts for high pollution levels.
- Allow users to report water contamination incidents for public awareness.

ACKNOWLEDGEMENT

We are expressing our sincere thanks to our correspondent **Adv.K.Ravindran Charles,** who is our guiding light and inspiration. We are greatful our principal **Dr.J.Japhynth, M.E.,Ph.D.,** for providing us the environment to develop this project. We are also thankful to **Dr.T.Jasperline, M.E., Ph.D.,** Head of the Department of Computer Science and Engineering for providing the necessary facilities during the execution of our project work. We also thank for her valuable suggestions, advice, guidance and constructive ideas in each and every step, which was indeed great need towards successful completion of the project. This project would not have been success without our internal guide. So, we would extend our deep sense of gratitude to our Internal guide **Mrs. A.Ruthrathnakumari M.E.,AP(CSE)** for excellent guidance and coordination, motivation and his efforts to bring out the best in us, his availability at all times and thought-provoking questions and timely suggestions. We are very much indebted to our department staff and students for relentlessy supporting us throughout our project work. Finally, our project would have been this shape without lovely efforts from our beloved parents. Theirinvaluable companionship, warmth, strong faith in our capabilities andthey have always helped us to be assertive in difficult times.

CONCLUSION

The Water Quality Regression project successfully demonstrates how machine learning can be leveraged to assess and predict water quality based on various physicochemical parameters. By implementing regression models, the system provides accurate predictions of the Water Quality Index (WQI), helping users evaluate the safety and usability of water sources. The project integrates Flask for backend processing and Streamlit for an interactive web interface, ensuring ease of use and accessibility. Users can input water quality parameters such as pH, TDS, BOD, COD, Turbidity, and DO, visualize data through interactive plots, and obtain real-time insights about water quality. The Water Quality Regression System represents a transformative advancement in environmental monitoring, successfully addressing the limitations of traditional water quality analysis methods through automation, integration, and accessibility.

By leveraging Streamlit and Python, the system unifies data preprocessing, regression modeling, interactive visualization, geospatial mapping, and real-time IoT integration into a single, user-friendly platform, eliminating the inefficiencies of fragmented workflows. Its ability to handle large datasets, train multiple regression models (e.g., Random Forest, SVR), and generate actionable insights through dynamic charts and maps empowers environmental scientists, researchers, and policymakers to make informed decisions for water resource management.

BIBLOGRAPHY

1. Tiwari, T. N., & Mishra, M. A. (1985). A new method for determining water quality index for rivers. International Journal of Environmental Studies, 26(3), 237-245.
2. Kumar, M., & Puri, A. (2012). A review of permissible limits of drinking water quality in India. Journal of Environmental Science & Engineering, 54(1), 94-100.
3. Sharma, S., & Bhardwaj, N. (2021). Machine learning-based water quality prediction models: A review. Environmental Monitoring and Assessment, 193(12), 784.
4. Garg, S., & Gupta, R. (2020). Real-time water quality monitoring and prediction using IoT and ML. Proceedings of the IEEE Conference on Smart Environments and Innovative Applications, 125-132.
5. Chaudhary, R., Singh, D. K., & Yadav, R. (2019). A comparative study of regression models for water quality prediction in Indian rivers. International Journal of Data Science and Analytics, 7(3), 192-205.
6. Government of India, National Water Mission (NWM). (2023). National Framework for Water Quality Management in India.

REFERENCE

https://www.bis.gov.in/

https://cpcb.nic.in/

https://jalshakti.gov.in/

https://www.who.int/water_sanitation_health