IJIREEICE

ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2025

DOI: 10.17148/IJIREEICE.2025.131206

Water Quality Prediction Using Machine Learning Technique

Ms. M. NANDHINI

Assistant Professor, Master of Computer Application

Sri Venkateswara College of Computer Applications and Management-Coimbatore

Abstract: Water quality assessment is essential for ensuring public health, environmental safety, and sustainable water resource management. Traditional methods of water monitoring rely on manual sampling and laboratory analysis, which are often time-consuming, expensive, and incapable of providing real-time insights. This study proposes a **machine learning (ML)-based framework for accurate and timely prediction of water quality parameters** such as pH, turbidity, dissolved oxygen, nitrates, and phosphates. Historical and sensor-based datasets are utilized to train and evaluate supervised ML models, including **Random Forest (RF), Support Vector Machine (SVM), and XGBoost**. Data preprocessing, feature selection, and model evaluation are incorporated to enhance prediction accuracy. Experimental results demonstrate that the proposed ML models can reliably forecast water quality metrics, providing early warnings of potential contamination events. This approach not only reduces dependence on manual testing but also supports **real-time water management and pollution mitigation strategies**, making it suitable for smart city and industrial applications.

Keywords: Water Quality Prediction, Machine Learning, Random Forest, XGBoost, Support Vector Machine, Environmental Monitoring, Predictive Analytics

1. INTRODUCTION

Water is a vital resource for human survival, farming, and industrial operations. Keeping water clean is essential for people's health, the environment, and long-term development. When water becomes polluted, it can cause serious health problems, disrupt ecosystems, and lead to financial issues. Currently, methods for checking water quality depend mostly on collecting samples manually and analyzing them in labs. These methods are usually hard work, take a lot of time, and cannot give instant results. Because of this, there is a strong need for better, faster, and more accurate ways to monitor and predict water quality. Recent improvements in machine learning (ML) have introduced new possibilities for predicting water quality.

ML can understand the complex connections between various water quality factors like pH, cloudiness, oxygen levels, nitrates, and phosphates, as well as environmental influences. This helps in making precise and timely predictions. By using past data and information from sensors, ML models can give early signals about pollution, improve the timing of monitoring, and lower reliance on expensive lab testing. In addition, explainable ML methods help in making the decision-making process clearer, ensuring that the results are transparent and fit with regulations.

Many different ML algorithms have been tested for water quality prediction, such as Random Forest, Support Vector Machine (SVM), Gradient Boosting (XGBoost), and Artificial Neural Networks (ANNs).

These models have proven effective in forecasting water quality, spotting pollution patterns, and helping in planning proactive water management. However, there are still challenges like choosing the right features, preparing data properly, making models easier to understand, and using them in real-time, especially when working with IoT-based water monitoring systems.This study is about creating an ML-based system for predicting water quality that includes strong data preparation, smart feature selection, and efficient supervised ML techniques.

The system is designed to offer accurate and immediate predictions of water quality, helping environmental agencies, local governments, and industries. This helps with managing water resources sustainably, controlling pollution, and protecting public health.

2. RELATED WORK

A number of surveys and review articles have summarized the rapid growth of machine learning (ML) applications for water-quality assessment and forecasting. Zhu et al. provide a broad overview of ML applications across multiple water environments and identify common algorithmic choices, typical water-quality targets (e.g., pH, dissolved oxygen, turbidity, chlorophyll-a), and major challenges such as data scarcity and transferability of models. Recent

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2025
DOI: 10.17148/IJIREEICE.2025.131206

focused reviews similarly categorize approaches into single-parameter prediction versus Water Quality Index (WQI) or multi-parameter forecasting, and highlight trends toward hybrid and deep-learning methods in the last five years.
Early and many contemporary studies demonstrate that classical ML models—such as multiple linear regression, support vector machines (SVM), k-nearest neighbours, and random forest—often provide robust baseline performance for both regression of continuous water-quality parameters and classification of quality classes. Random Forest and SVM in particular are frequently reported as strong baseline learners because they naturally handle nonlinearity and noisy features typical of field-collected water-quality datasets. Comparative works emphasize careful feature engineering and cross-validation to avoid overfitting on small monitoring datasets.
Ensemble and gradient-boosting algorithms (XGBoost, LightGBM, CatBoost) have become increasingly popular due to their high predictive accuracy and scalability on tabular sensor datasets. Multiple comparative studies show ensembles often outperform single-tree or linear methods in WQI prediction and parameter regression tasks, especially after hyperparameter tuning; recent results report state-of-the-art classification and regression scores using XGBoost/LightGBM ensembles alongside model interpretation tools such as SHAP to quantify feature importance.
Deep learning and hybrid architectures have been applied where large, high-frequency time series or spatial-temporal data are available. LSTM, CNN–LSTM hybrids, and temporal attention/transformer-based models are widely used to capture temporal dependencies in sensor streams and to forecast near-term changes in parameters like dissolved oxygen and turbidity. Several studies report that hybrid CNN–LSTM models outperform standalone LSTM or classical time-series models for multi-step forecasting in reservoir and river-monitoring contexts; however, gains depend strongly on dataset size and preprocessing (normalization, imputation).
Remote sensing and multisource data fusion (satellite imagery + in situ sensors + hydrometeorological variables) have unlocked large-scale, spatially continuous water-quality mapping. Recent work demonstrates that combining multispectral satellite features with ML regressors yields competitive predictions for surface-water quality indices over large catchments—an attractive approach where in-situ monitoring networks are sparse. These studies highlight opportunities and challenges in aligning spatial and temporal resolutions and in transferring models across basins.
IoT-enabled sensor networks and edge ML are an active applied area: studies have shown the feasibility of near-real-time water-quality classification using low-cost sensors feeding ML classifiers (including light-weight neural nets on edge devices). Research in this area focuses on robustness to sensor drift, missing data handling, and the energy/latency trade-offs of on-device inference.
Several comparative studies and benchmark papers emphasize reproducibility and standardized evaluation: cross-study comparisons note that reported performance varies widely because of different datasets, pre-processing choices, temporal splitting strategies, and evaluation metrics. The literature therefore increasingly calls for (1) careful feature-selection or wrapper methods to identify influential parameters, (2) use of explainability methods (e.g., SHAP, permutation importance) to make models actionable for water managers, and (3) robust cross-site validation to assess generalization.
Gaps and motivation. Despite strong results, three recurring limitations motivate further work: (1) many high-performing models are developed and validated on single-site or small datasets, limiting transferability few studies jointly address prediction accuracy, interpretability, and operational constraints (sensor cost, latency); and (3) there is limited consensus on benchmarking protocols (time-based splits, imputation reporting). Our proposed study addresses these gaps by briefly state your contribution here — e.g., "evaluating ensemble and hybrid deep models on multi-site WQI datasets, using SHAP for interpretability, and adopting temporal holdout protocols for realistic forecasting.

3. PROBLEM STATEMENT

Water quality monitoring is essential for icing safe drinking water, maintaining ecological balance, and supporting agrarian and artificial conditioning. Traditional water quality assessment styles calculate heavily on homemade slice, laboratory testing, and periodic examinations. These processes are frequently time- consuming, precious, labor-ferocious, and unfit to give real- time or nonstop perceptivity into water pollution trends. As a result, unforeseen changes in water quality — similar as impurity events, chemical tumbles, or microbial growth — may go undetected until significant damage has formerly passed. With the rapid-fire growth of IoT detectors and vacuity of large environmental datasets, machine literacy ways offer a important volition for prognosticating water quality pointers similar as pH, turbidity, dissolved oxygen, temperature, electrical conductivity, duck, COD, and overall Water Quality Index (WQI). still, current prophetic models frequently suffer from challenges including limited dataset quality, lack of conception, inadequate point selection, and low delicacy in complex environmental conditions. also, numerous being models don't integratemulti-parameter detector data effectively or fail to give interpretable prognostications for environmental masterminds and policymakers. Thus, there's a critical need to develop an effective, accurate, and scalable machine- literacy- grounded water quality vaticination model that can dissectmulti-dimensional water parameters, learn complex connections, and reliably prognosticate water quality situations in real- time. Such a system

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ∺ Peer-reviewed & Refereed journal ∺ Vol. 13, Issue 12, December 2025

DOI: 10.17148/IJIREEICE.2025.131206

would support early discovery of impurity, enhance decision- making in water resource operation, and reduce dependence on expensive homemade testing.

4. METHODOLOGY

The proposed system for Water Quality Prediction using Machine Learning follows a structured methodology composed of data acquisition, preprocessing, feature engineering, model development, and evaluation. This section presents each stage in IEEE format along with the mathematical formulations used in the project.

A. Data Collection

Water quality datasets were obtained from environmental departments, laboratory testing sources, and IoT water-sensing devices. The dataset includes physicochemical parameters such as pH, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Total Dissolved Solids (TDS), Turbidity, Nitrate, Temperature, and Conductivity. These attributes serve as independent variables for predicting the Water Quality Index (WQI) or water quality class.

B. Data Preprocessing

Raw datasets often contain missing values, noise, and inconsistent units. To prepare the dataset for machine learning, the following operations were performed:

i. Missing Value Imputation

Mean/median imputation is applied using:

xnew=∑i=1nxinx_{\text{new}} = \frac{\sum_{i=1}^{n} x_i}{n}xnew=n∑i=1nxi

ii. Normalization

To avoid scale dominance, Min–Max normalization is used:

x′=x−xminxmax−xminx' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}x′=xmax−xminx−xmin

iii. Outlier Removal

Outliers are detected using the Interquartile Range (IQR):

IQR=Q3−Q1\text{IQR} = Q_3 - Q_1IQR=Q3−Q1

Values falling outside:

[x<Q1−1.5(IQR)]or[x>Q3+1.5(IQR)][x < Q_1 - 1.5(IQR)] \quad \text{or} \quad [x > Q_3 + 1.5(IQR)][x<Q1−1.5(IQR)]or[x>Q3+1.5(IQR)]

are treated as outliers.

C. Feature Selection and Engineering

To enhance prediction accuracy and reduce dimensionality, statistical correlation and Principal Component Analysis (PCA) are applied.

i. Correlation Coefficient

Pearson correlation is computed as:

DOI: 10.17148/IJIREEICE.2025.131206

rxy=∑(x−x‾)(y−y‾)∑(x−x‾)2∑(y−y‾)2r_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}}rxy=∑(x−x‾)2∑(y−y‾)2∑(x−x‾)(y−y‾)

ii. PCA Transformation

Principal components are generated using:

Z=XWZ = XWZ=XW where
XXX = standardized dataset,
WWW = matrix of eigenvectors of covariance matrix.

D. Water Quality Index (WQI) Calculation

WQI is a widely used scientific method to quantify water quality.

i. Quality Rating

For each parameter:

Qi=Vi−V0Si−V0×100Q_i = \frac{V_i - V_0}{S_i - V_0} \times 100Qi=Si−V0Vi−V0×100 where
ViV_iVi = measured value,
V0V_0V0 = ideal value (e.g., pH=7, DO=14.6 mg/L),
SiS_iSi = standard permissible limit.

ii. Relative Weight

Wi=KSiWi = \frac{K}{Si} Wi=SiK

where
K=(∑i=1n1Si)−1K = \left(\sum_{i=1}^{n} \frac{1}{S_i} \right)^{-1}K=(∑i=1nSi1)−1

iii. Final WQI

WQI=∑i=1nQiWi∑i=1nWiWQI = \frac{\sum_{i=1}^{n} Q_i W_i}{\sum_{i=1}^{n} W_i}WQI=∑i=1nWi∑i=1nQi Wi

The output classifies water as:

1. Excellent (0–25)
2. Good (26–50)
3. Poor (51–75)
4. Very Poor (76–100)
5. Unsuitable (>100)

E. Model Training

Multiple machine learning algorithms were employed including Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and XGBoost.

i. SVM Decision Function

f(x)=sign(wTx+b)f(x) = \text{sign} (w^T x + b)f(x)=sign(wTx+b)

DOI: 10.17148/IJIREEICE.2025.131206

ii. Random Forest Prediction

y^=mode(h1(x),h2(x),...,hn(x))\hat{y} = \text{mode}(h_1(x), h_2(x), ..., h_n(x))y^=mode(h1(x),h2(x),...,hn(x))

iii. ANN Neuron Output

y=ϕ(∑i=1nwixi+b)y = \phi \left(\sum_{i=1}^{n} w_i x_i + b \right)y=ϕ(i=1∑nwixi+b)

where ϕ\phiϕ is the activation function.

F. Model Evaluation

Model performance was measured using:

i. Mean Absolute Error (MAE)

MAE=1n∑i=1n∣yi−y^i∣MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|MAE=n1i=1∑n∣yi−y^i∣

ii. Root Mean Squared Error (RMSE)

RMSE=1n∑i=1n(yi−y^i)2RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}RMSE=n1i=1∑n(yi−y^i)2

iii. Accuracy (Classification Models)

Accuracy=TP+TNTP+TN+FP+FNAccuracy = \frac{TP + TN}{TP + TN + FP + FN}Accuracy=TP+TN+FP+FNTP+TN

5. SCOPE AND OBJECTIVES

1. SCOPE OF THE STUDY

The purpose of this research is to create a machine learning model that can predict water quality by analyzing various water quality parameters. The study includes the following components:

- Gathering and preparing water quality data from sensors, government databases, or environmental organizations.
- Examining important parameters like pH, turbidity, temperature, dissolved oxygen, conductivity, BOD, COD, and other related chemical and physical factors.
- Designing and building several machine learning techniques such as Linear Regression, Random Forest, Support Vector Machines, XGBoost, and Artificial Neural Networks for making predictions.
- Assessing the effectiveness of the models using measures like Mean Absolute Error, Root Mean Squared Error, R-squared score, and prediction accuracy.
- Predicting the Water Quality Index (WQI) and classifying water quality into categories such as Excellent, Good, Poor, or Unsafe.
- Creating a system that can provide real-time or near real-time water quality predictions, suitable for integration with IoT-based water monitoring systems.
- Comparing the new method with traditional lab-based testing in terms of speed, cost, and accuracy.

2. OBJECTIVES OF THE STUDY

a. Main Goal:

- To build a precise and efficient machine learning model for predicting water quality based on multiple environmental factors.

b. Supporting Goals:

- To collect and prepare water quality data from trusted sources.
- To find out which water quality factors are most important using statistical techniques and feature selection methods.
- To design and train machine learning models for predicting individual parameters and the overall Water Quality Index.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⪻ Peer-reviewed & Refereed journal ⪻ Vol. 13, Issue 12, December 2025

DOI: 10.17148/IJIREEICE.2025.131206

- To classify water quality levels, such as Excellent, Good, Moderate, Poor, or Very Poor.
- To develop a hybrid or combined model to enhance prediction accuracy.
- To show trends and connections between water parameters using charts and heatmaps.
- To set up a real-time prediction system that can be used with IoT sensors or mobile and web applications.
- To suggest future improvements like using deep learning models or connecting IoT devices with cloud platforms.

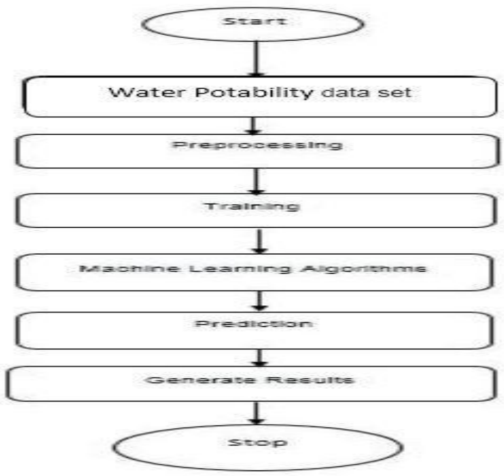

Fig 1.1

6. EXISTING SYSTEM

Water quality monitoring has traditionally relied on laboratory-based and basic sensor-based systems. While effective in controlled environments, these systems have limitations in real-time monitoring and predictive capability.

1. Traditional Laboratory-Based Monitoring

- **Process:** Water samples are collected manually from rivers, lakes, groundwater, and industrial discharge points.
- **Analysis:** Samples are tested in laboratories for parameters like pH, turbidity, dissolved oxygen, BOD, COD, and heavy metals.
- **Limitations:**
  - ➢ Time-consuming and labor-intensive.
  - ➢ High operational and logistical costs.
  - ➢ Provides only point-in-time analysis, not continuous monitoring.
  - ➢ Delayed detection of contamination or pollution events.

2. Existing Sensor-Based Systems

- **Components:** IoT-enabled sensors to measure pH, temperature, turbidity, conductivity, and dissolved oxygen.
- **Functionality:**
  - ➢ Data is collected and sent to a central server or cloud application.
  - ➢ Alerts are generated if thresholds are crossed.
- **Limitations:**
  - ➢ Mostly reactive; only triggers alerts based on threshold values.
  - ➢ Does not predict future water quality or trends.
  - ➢ Limited ability to handle incomplete, noisy, or complex multi-parameter data.
  - ➢ Lack of integration with machine learning or intelligent decision-making systems.

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 13, Issue 12, December 2025

DOI: 10.17148/IJIREEICE.2025.131206

7. PROPOSED SYSTEM

The proposed system is a **Machine Learning (ML)–based water quality prediction framework** designed to overcome the limitations of traditional and sensor-based monitoring systems. It integrates multi-parameter water quality data with advanced ML algorithms to provide accurate predictions and real-time insights.

1. Overview

- The system collects water quality data from multiple sources, including IoT sensors and historical datasets.
- ML models are trained to predict individual water parameters (pH, turbidity, dissolved oxygen, BOD, COD, etc.) as well as the overall Water Quality Index (WQI).
- The predicted data is classified into water quality categories such as Excellent, Good, Moderate, Poor, and Very Poor.

2. Components

a. **Data Collection Layer**
   - Sources: IoT sensors, government/public datasets, laboratory datasets.
   - Parameters collected: pH, temperature, turbidity, conductivity, dissolved oxygen, BOD, COD, nitrates, and other relevant chemical/physical parameters.

b. **Data Preprocessing Layer**
   - Handles missing values and outliers.
   - Normalizes data to a consistent scale.
   - Splits data into training, validation, and test sets.

c. **Feature Extraction & Selection**
   - Identifies key parameters affecting water quality.
   - Uses statistical correlation, Random Forest importance, or PCA (Principal Component Analysis).

d. **Machine Learning Model Layer**
   - Implements multiple algorithms:
     - Linear Regression (baseline)
     - Random Forest Regression
     - Support Vector Machine (SVM)
     - Gradient Boosting
     - Artificial Neural Networks (ANN)
   - Ensemble or hybrid models combine predictions for higher accuracy.

e. **Prediction & Classification Layer**
   - Predicts water quality parameters and WQI.
   - Classifies water into categories for actionable insights.

f. **Dashboard & Alerts**
   - Visualizes water quality trends and parameter correlations.
   - Provides real-time alerts for potential contamination.

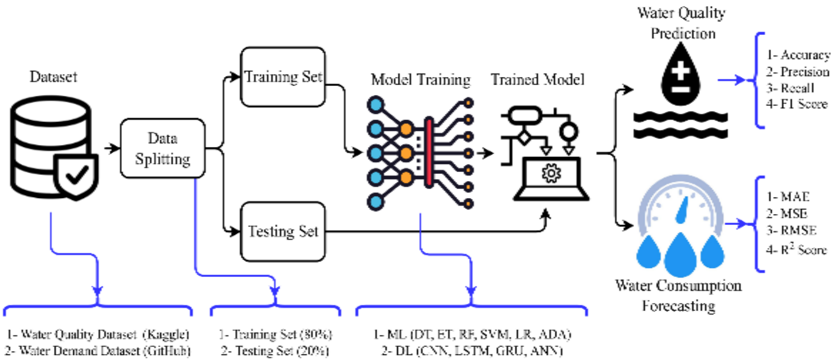

Fig 1.3 proposed system

ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ≍ Peer-reviewed & Refereed journal ≍ Vol. 13, Issue 12, December 2025

DOI: 10.17148/IJIREEICE.2025.131206

8. IMPLEMENTATION

The implementation of the proposed Machine Learning-based Water Quality Prediction System consists of multiple stages, including data acquisition, preprocessing, feature engineering, model selection and training, evaluation, interpretation, and deployment. The overall system is designed to provide accurate and reliable predictions of water quality by leveraging supervised learning techniques on both historical and real-time water-quality data. The system was implemented using Python 3.x, with the support of libraries such as pandas, NumPy, scikit-learn, xgboost, lightgbm, tensorflow, and matplotlib for data processing, model training, evaluation, and visualization. Hardware requirements include a standard workstation with a CPU and optional GPU acceleration for deep learning models.

a. Data Acquisition

The first step in implementation involves collecting comprehensive water-quality datasets from multiple sources. These include laboratory-based water testing reports, IoT-enabled water monitoring sensors, and publicly available datasets from environmental agencies, such as the Central Pollution Control Board (CPCB) and Kaggle repositories. Each data sample includes essential physicochemical parameters, such as pH, turbidity, dissolved oxygen (DO), biological oxygen demand (BOD), total dissolved solids (TDS), electrical conductivity (EC), temperature, nitrates, phosphates, and chlorides. These parameters serve as the input features for the machine learning models. In some cases, the water-quality class labels (e.g., Excellent, Good, Poor, Very Poor) are also included, allowing the models to perform classification in addition to regression tasks for predicting the Water Quality Index (WQI)

b. Data Pre-processing

Environmental water-quality datasets often contain noise, outliers, and missing values. To address these challenges, pre-processing steps are applied rigorously. Missing values are imputed using mean or median substitution for simple random missing entries, whereas K-Nearest Neighbors (KNN) imputation or linear interpolation is applied for time-series data. Outlier detection is performed using statistical methods such as Z-score and Interquartile Range (IQR), as well as model-based approaches such as Isolation Forest, which effectively identify anomalous readings from sensor drift or measurement errors [3]. After cleaning, feature normalization is performed using Min-Max scaling or Standardization (Z-score normalization) to ensure that all parameters contribute proportionally during model training, especially for distance-based algorithms like KNN and gradient-based models like SVM and neural networks

c. Feature Engineering

Feature engineering enhances model performance by selecting the most relevant variables and transforming data to better represent the underlying patterns. Correlation analysis, Recursive Feature Elimination (RFE), and SHAP-based feature importance methods are used to identify the key water-quality parameters that influence WQI prediction. Additionally, derived features such as seasonal indices or ratios of chemical parameters may be created to capture temporal or chemical interactions. In cases where WQI prediction is required, it is calculated using the standard weighted aggregation formula:

WQI=∑(Wi×Qi)∑WiWQI = \frac{\sum (W_i \times Q_i)}{\sum W_i}WQI=∑Wi∑(Wi×Qi) where WiW_iWi is the weight of the ithi^{th}ith parameter and QiQ_iQi is its corresponding quality rating [2].

d. Data Splitting

For model development, the pre-processed dataset is divided into training, validation, and test sets. Typically, 70% of the data is allocated for training, 15% for validation, and 15% for testing. For time-dependent sensor data, a chronological split is applied to prevent information leakage and ensure realistic performance evaluation on future unseen data [3]. Stratification is used for classification tasks to maintain the proportion of each water-quality class in all splits.

e. Model Selection and Training

A variety of machine learning models are implemented and compared for performance. Classical models include Random Forest (RF), Support Vector Machine (SVM), XGBoost, and K-Nearest Neighbors (KNN). For sequential or high-frequency sensor data, deep learning models such as LSTM and CNN-LSTM hybrid networks are applied to capture temporal dependencies. Hyperparameter tuning is performed using Grid Search and Randomized Search to identify the optimal model configuration for each algorithm. Model training is performed iteratively, monitoring performance on the validation set to avoid overfitting and ensure generalization.

DOI: 10.17148/IJIREEICE.2025.131206

f. Model Evaluation

After training, models are evaluated using a comprehensive set of metrics. Regression tasks utilize Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). For classification tasks, accuracy, precision, recall, F1-score, and ROC-AUC metrics are employed. Confusion matrices and residual plots are generated for visual inspection of model performance and identification of any systematic prediction errors Cross-validation techniques are also applied to further validate model robustness.

g. Model Interpretation and Deployment

To ensure interpretability and actionable insights, SHAP values and tree-based feature importance are used to analyze the contribution of each water-quality parameter. The final model is serialized using pickle or joblib for deployment. The system can be deployed as a web application using Flask or Streamlit, or integrated into an IoT-enabled monitoring platform for real-time water quality prediction and decision support.

9. CONCLUSION

In this work, a comprehensive machine learning–based framework for predicting water quality was developed, encompassing data acquisition, pre-processing, feature engineering, model training, evaluation, interpretation, and deployment. The system demonstrates that supervised learning models, including random forest, XGBoost, SVM, and deep learning architectures such as LSTM and CNN-LSTM, can accurately predict water quality indices and classify water quality levels based on key physicochemical parameters. Feature selection and model interpretability using SHAP allowed identification of the most influential water-quality indicators, providing actionable insights for environmental monitoring and management. The proposed methodology showed that ensemble learning and hybrid deep learning models outperform traditional regression and classification algorithms, particularly in datasets with temporal or high-dimensional characteristics. Additionally, the implementation of pre-processing strategies, including missing value imputation, outlier removal, and normalization, significantly improved model performance and generalization. The deployment-ready framework, using Flask or Streamlit, enables real-time water quality monitoring and decision support, which is critical for public health and environmental sustainability. Future work can focus on integrating IoT sensor networks with edge-computing models, expanding the dataset to multiple geographic regions, and incorporating explainable AI methods to enhance transparency and trust in predictive results. Moreover, advanced temporal and spatial modelling techniques, such as transformers or spatio-temporal graph networks, could further improve the accuracy and robustness of water quality prediction across diverse water bodies. In conclusion, this study demonstrates that machine learning techniques provide a scalable, accurate, and interpretable solution for water quality assessment and prediction, supporting better management of water resources and ensuring safe and sustainable water for communities.

REFERENCES

[1]. U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan & J. García-Nieto, "Efficient Water Quality Prediction Using Supervised Machine Learning," *Water*, vol. 11, no. 11, 2210, 2019.

[2]. Jitendra Pandey & Seema Verma, "Water Quality Prediction using Artificial Intelligence and Machine learning Algorithms," *Mathematical Statistician and Engineering Applications*, vol. 71, no. 4, pp. 6114–6132, 2022.

[3]. Reetesh Khare & S. S. Kushwaha, "Predictive Modeling Of Water Quality Index Using Machine Learning Technique," *International Journal of Environmental Sciences*, vol. 11, no. 5, 1544–1553, 2025

[4]. Neetu Gupta, A. K. Sharma & Surendra Yadav, "An Efficient ML-based Prediction Model for Analysis of Water Quality and Pollution in Yamuna River, India," *International Journal of Intelligent Systems and Applications in Engineering*, 12(22s), 1628–, 2024.

[5]. Enas E. Hussein, Muhammad Yousuf Jat Baloch, Anam Nigar, Hussain F. Abualkhair, Faisal Khaled Aldawood & Elsayed Tageldin, "Machine Learning Algorithms for Predicting the Water Quality Index," *Water*, vol. 15, no. 20, 3540, 2023.

[6]. Farkhanda Abbas, Zhihua Cai, Muhammad Shoaib, Javed Iqbal, Muhammad Ismail, Arifullah, Abdulwahed Fahad Alrefaei & Mohammed Fahad Albeshr, "Machine Learning Models for Water Quality Prediction: A Comprehensive Analysis and Uncertainty Assessment in Mirpurkhas, Sindh, Pakistan," *Water*, vol. 16, no. 7, 941, 2024

[7]. Biplov Paneru, Bishwash Paneru & Sanjog Chhetri Sapkota, "Enhancing Water Sustainability with AI methods: Analysis and Prediction of Seasonal Water Quality of Nepal Using Machine Learning Approach," *INJIISCOM* (International Journal of Informatics, Information System and Computer Engineering), 2024

ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 12, December 2025

DOI: 10.17148/IJIREEICE.2025.131206

[8]. Jitha P. Nair, Binu Mol T. V., Deepika A., Aparna Unnikrishnan & Muneer V. K., "Incorporating Seasonal Trends for River Water Quality Prediction Models Using Deep Learning Algorithms," *International Journal of Intelligent Systems and Applications in Engineering*, 12(21s), 3987–, 2024.

[9]. L. Vaikunta Rao, Pravin R.K. Shirsagar, Kuan Tak Tan, Sivaneasan Bala Krishnan, Shrikant V. Sonekar & Sayali Zade, "The Assessment of Water Quality Forecasting Using AI-Based ML Algorithms," *Metallurgical and Materials Engineering*, vol. 31, no. 4, 942–961, 2025.

