Comparative Study on Water Potability Prediction using Ensembled Based Techniques

Akshatha M. R.*, Chitra K.

Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India.

***Email:** akshathamr0704@gmail.com

Abstract

Water quality assessment plays a vital role in public health protection and environmental sustainability. Conventional testing techniques, though accurate, are time-consuming, labour-intensive, and prone to human error. Recent advancements in Machine Learning (ML), Deep Learning (DL), and the Internet of Things (IoT) have transformed water potability prediction through intelligent, automated systems. This paper presents a comparative review of ensemble and hybrid ML/DL approaches such as Bagging, Gradient Boosting, XGBoost, and stacked models that have achieved accuracies ranging from 83% to 99.6% in recent studies between 2023 to 2025. Furthermore, IoT-based sensors and blockchain integration enable real-time monitoring, transparency, and data security in water management frameworks. This work highlights current trends, research gaps, and emerging innovations focusing on adaptive, scalable, and secure water quality prediction systems for sustainable smart water management.

Keywords

Water potability, Machine learning, Ensemble methods, IoT, Blockchain

Introduction

Access to clean drinking water is essential for human health, yet maintaining consistent water quality remains a global challenge. Traditional lab-based testing is slow, costly, and unsuitable for large-scale or real-time monitoring. With the growth of water quality datasets, Machine Learning (ML) and Deep Learning (DL) techniques have emerged as powerful tools for automated potability prediction. Ensemble models like Bagging, Gradient Boosting, and XGBoost achieve accuracies above 95%, while deep networks such as CNNs and DNNs capture complex non-linear patterns. IoT devices enable real-time sensing, and blockchain improves data security and traceability.

However, current models still face issues like limited dataset diversity, poor adaptability to changing environmental conditions, and weak integration with secure frameworks. This study provides a comparative review of ensemble and hybrid ML/DL approaches for water potability prediction and examines the role of IoT and blockchain in intelligent water monitoring. The goal is to assess model performance, highlight emerging

Submission: 28 September 2025; **Acceptance:** 11 December 2025; **Available online:** December 2025

trends, and propose future directions for adaptive, scalable, and secure water quality assessment systems.

Danush et al. (2023) analyzed Korattur Lake water using XGBoost, KNN, Bagging, and AdaBoost, where Bagging with Decision Trees achieved the best accuracy (99.67%) and high speed, identifying pH, TDS, and turbidity as key predictors. Meenalochani et al. (2024) built an IoT edge system with Raspberry Pi and sensors, where the Decision Tree classifier (83%) outperformed RF and Logistic Regression, integrating results into a mobile app via Twilio API. Tummala et al. (2024) proposed a stacked hybrid Gradient Boosting–Ridge Regression model for Water Quality Index prediction ($R^2 = 0.9899$), suggesting deep learning integration. Shibin et al. (2024) applied the GEM continual learning algorithm for potability prediction (87.2%), surpassing RF and LGBM. Huynh et al. (2024) used CNN, LSTM, GRU, and TCN for river salinity forecasting, with CNN yielding the lowest error (1.40 mg/L).

Han et al. (2024) predicted boiler drum water levels using Random Forest with high accuracy under variable conditions. Dumbre et al. (2024) used Gradient Boosting, Neural Networks, and Decision Trees for water safety prediction, achieving up to 0.998 accuracy. Rao et al. (2024) combined RBM and ANN, attaining 69% accuracy and emphasizing larger datasets. Chaudhari et al. (2025) classified potable vs. non-potable water using SVM and IoT sensors, recommending suitable filters. Nair et al. (2025) developed a DNN-based water safety framework (92%) integrated with blockchain for secure storage and smart alerts.

Methodology

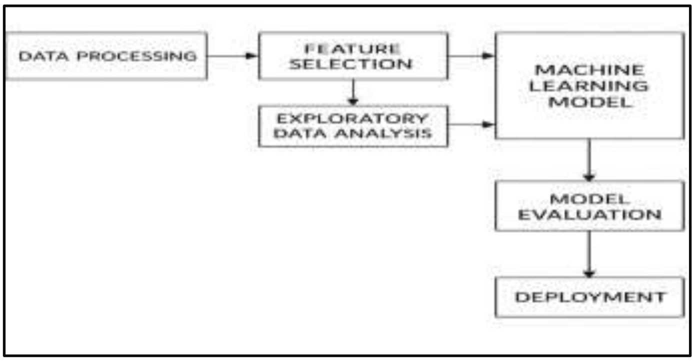
The studies examined between 2023 and 2025 adopted diverse methodologies for predicting water potability, monitoring water quality, and forecasting environmental parameters. The approaches integrate machine learning (ML), deep learning (DL), IoT-enabled sensing, ensemble techniques, and blockchain frameworks. Despite their diversity, most works follow a common layered methodology that includes data acquisition, pre-processing, predictive modeling, system integration, and evaluation.

Figure 1. System Architecture

A. Data Acquisition and Preprocessing

Data acquisition methods vary depending on the application domain.

- **Sensor-based acquisition:** Studies such as Meenalochani et al. (2024) and Chaudhari et al. (2025) utilized IoT-enabled sensors (pH, TDS, turbidity, conductivity) connected to Raspberry Pi devices for real-time water monitoring.



- **Public datasets:** Rao et al. (2024), Danush et al. (2023), and Nair et al. (2025) employed Kaggle and UCI benchmark datasets for training and validation of predictive models.

Preprocessing techniques commonly included median imputation for handling missing data, normalization for feature scaling, outlier detection, and bootstrapping for data balancing. Feature importance analysis (e.g., using Random Forest and Gradient Boosting) identified key predictors such as pH, TDS, and turbidity as dominant parameters influencing water potability.

B. Predictive Modeling Approaches

A wide range of models were implemented across studies depending on prediction objectives:

- **Classical ML models:** Decision Trees, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) (Chaudhari et al., Han et al., Meenalochani et al.).
- **Ensemble methods:** Bagging, AdaBoost, Gradient Boosting, and XGBoost (Danush et al., Dumbre et al., Tummala et al.), which consistently outperformed individual classifiers.
- **Hybrid and stacked architectures:** Tummala et al. (2024) developed a stacked model combining Gradient Boosting with Ridge Regression, while Rao et al. (2024) used Restricted Boltzmann Machines (RBM) for feature extraction followed by Artificial Neural Networks (ANN) for prediction.
- **Deep learning frameworks:** Huynh et al. (2024) applied CNN, LSTM, GRU, and TCN for salinity forecasting, while Nair et al. (2025) employed Deep Neural Networks (DNN) for water potability prediction.
- **Continual learning:** Shibin et al. (2024) introduced the Gradient Episodic Memory (GEM) algorithm to enable adaptive learning in dynamic water quality environments.

C. System-Level Integration

Several studies emphasized integrating predictive models into complete water monitoring systems:

- **IoT frameworks:** Chaudhari et al. (2025) and Meenalochani et al. (2024) combined IoT sensors with edge computing and mobile-based alert systems.
- **Blockchain solutions:** Nair et al. (2025) implemented Hyperledger Fabric for secure, tamper-proof data storage and smart contract-based automation.
- **Industrial applications:** Han et al. (2024) used Random Forest for real-time prediction of boiler drum water levels under varying industrial load conditions.
- **Edge computing:** Meenalochani et al. (2024) utilized Raspberry Pi devices and Twilio API for generating live contamination alerts.

D. Evaluation Metrics

Performance evaluation across studies was conducted using a range of statistical metrics, including accuracy, precision, recall, F1-score for classification, and R^2, mean error, or execution time for regression and forecasting tasks.

- Danush et al. (2023): Bagging achieved 99.67% accuracy; XGBoost reached 99.2%.
- Tummala et al. (2024): Stacked hybrid model achieved $R^2 = 0.9899$.
- Huynh et al. (2024): CNN achieved the lowest error (1.40 mg/L) for 1-day salinity forecasting.
- Shibin et al. (2024): GEM achieved 87.2% accuracy, outperforming traditional Random Forest and LGBM models.

E. Emerging Methodological Trends

Recent research highlights several promising trends in water potability prediction:

- Hybrid modeling combining deep learning and statistical forecasting (Huynh et al., 2024).
- Adaptive continual learning for dynamic water quality assessment (Shibin et al., 2024).
- Blockchain-based secure monitoring using smart contracts (Nair et al., 2025).
- IoT–satellite data fusion for large-scale environmental monitoring (Nair et al., 2025).
- Edge and cloud-based deployment for real-time prediction and alert systems (Meenalochani et al., 2024).

Results and Discussion

The studies reviewed between 2023 and 2025 show significant advancements in Machine Learning (ML), Deep Learning (DL), and IoT-based systems for water quality prediction. Research clearly reflects a shift from conventional ML models toward hybrid ensemble frameworks and deep learning architectures, often combined with IoT and blockchain for real-time monitoring, scalability, and secure data handling. This evolution indicates a growing focus not only on improving predictive accuracy but also on deploying intelligent, adaptable systems suitable for practical environments.

Traditional ML models continue to deliver strong and reliable performance. Danush et al. (2023) achieved 99.67% accuracy using Bagging on the Korattur Lake dataset, slightly outperforming XGBoost (99.2%). Han et al. (2024) used Random Forest for industrial boiler water level prediction, maintaining stable accuracy under fluctuating conditions. Chaudhari et al. (2025) applied SVM with IoT sensor data and achieved high accuracy in classifying potable and non-potable water. These findings show that traditional ML models remain efficient, interpretable, and computationally lightweight, making them suitable for structured datasets and real-time applications.

Hybrid and ensemble approaches demonstrated even higher generalization capabilities. Tummala et al. (2024) achieved an R^2 of 0.9899 using a stacked Gradient Boosting–Ridge Regression model, while Dumbre et al. (2024) reported accuracies up to 0.998 using Gradient Boosting and Neural Networks. Shibin et al. (2024) introduced the GEM continual learning model, reaching 87.2% accuracy and outperforming Random Forest and LGBM baselines. These improvements stem from model diversity, which reduces overfitting, enhances stability, and improves predictive consistency across varied datasets.

Deep learning further expands predictive capabilities by capturing complex non-linear and temporal patterns. Huynh et al. (2024) evaluated CNN, LSTM, GRU, and TCN models, with CNN achieving the lowest error of 1.40 mg/L in salinity forecasting. Rao et al. (2024) combined RBM with ANN, whereas Nair et al. (2025) achieved 92% accuracy using DNN integrated with blockchain to ensure data security. While DL models provide better adaptability, they require large datasets and higher computational resources, which may limit their feasibility in low-resource settings.

IoT and blockchain integration has strengthened real-time and secure water quality monitoring. Meenalochani et al. (2024) used Raspberry Pi sensors and Twilio API to achieve 83% accuracy with real-time alerts, while Chaudhari et al. (2025) combined IoT data with ML models for potability detection and filter recommendations. Nair et al. (2025) used Hyperledger

Fabric to ensure tamper-proof data storage. These integrations highlight the shift from standalone prediction models to complete intelligent systems with improved usability, transparency, and trustworthiness.

Overall, the reviewed studies shows a clear trend toward more adaptive, scalable, and secure water monitoring solutions. Ensemble methods provide stable and accurate results through model diversity, while CNN and other DL models excel at learning complex patterns. The integration of IoT and blockchain bridges the gap between theoretical prediction and practical deployment, paving the way for intelligent and sustainable water quality management systems. Table 1 shows the comparison of reviewed studies between 2023 to 2025.

Table 1. Comparison of Reviewed Studies (2023–2025)

Ref.	Author(s), Year	Method(s) Used	Best Performance / Findings
[1]	Danush et al., 2023	Bagging, XGBoost, KNN, AdaBoost	Bagging achieved 99.67% accuracy, XGBoost 99.2%; identified pH, TDS, turbidity as key predictors.
[2]	Meenalochani et al., 2024	Decision Tree, Random Forest, Logistic Regression (IoT + Edge Computing)	Decision Tree achieved 83% accuracy; implemented Raspberry Pi + Twilio API for real-time alerts.
[3]	Tummala et al., 2024	Stacked Gradient Boosting + Ridge Regression	Achieved $R^2 = 0.9899$; demonstrated improved accuracy for WQI forecasting.
[4]	Shibin et al., 2024	GEM Continual Learning, Random Forest, LGBM	Achieved 87.2% accuracy, outperforming RF (68.1%) and LGBM (66.9%).
[5]	Huynh et al., 2024	CNN, LSTM, GRU, TCN, ARIMA, ANN	CNN achieved 1.40 mg/L error for 1-day prediction, outperforming others.
[6]	Han et al., 2024	Random Forest (WEKA)	High accuracy and low error for boiler drum water level prediction.
[7]	Dumbre et al., 2024	Gradient Boosting, Neural Networks, Decision Trees	Ensemble achieved 0.998 accuracy using Orange ML platform.
[8]	Rao et al., 2024	RBM + ANN	Achieved 69% accuracy, stable results via bootstrapping.
[9]	Chaudhari et al., 2025	SVM, IoT Integration	SVM achieved best performance; system recommended RO or activated carbon filters.
[10]	Nair et al., 2025	DNN, XGBoost, SVM + Blockchain (Hyperledger Fabric)	DNN achieved 92% accuracy; blockchain ensured secure and transparent data storage.

Conclusion

The review of research conducted between 2023 to 2025 demonstrates the remarkable progress in applying machine learning, deep learning, and IoT-driven approaches to water quality prediction and monitoring. Traditional ML models such as Bagging and SVM have achieved near-perfect accuracy in classifying potable versus non-potable water, while hybrid and ensemble methods, including stacked regressors and gradient boosting, have proven highly effective for Water Quality Index forecasting. Deep learning architectures, particularly CNNs, have shown superior performance in time-series forecasting tasks such as river salinity prediction.Beyond predictive accuracy, recent studies have placed greater emphasis on practical deployment and scalability**.** IoT-enabled frameworks now allow real-time monitoring with mobile alerts, and blockchain has been integrated to enhance data security and trust. These innovations highlight a clear shift from purely computational models toward end-to-end intelligent systems that combine prediction, monitoring, and decision support.

Common findings across the reviewed works indicate that pH, TDS, turbidity, conductivity, and sulfate remain the most critical parameters for predicting water quality, reinforcing consensus between data-driven methods and expert knowledge. Importantly, continual learning models and IoT–blockchain integrations represent the next frontier, addressing the challenges of adaptivity, reliability, and secure information sharing.However, limitations such as dataset imbalance, high computational costs, and lack of large-scale deployment remain challenges. Future research should focus on lightweight, adaptive models integrated with IoT–blockchain systems for real-time global water monitoring.In conclusion, the integration of ML/DL algorithms with IoT and blockchain technologies has the potential to transform water quality management into a smart, adaptive, and reliable system**.**

Acknowledgement

There are no grant or funding bodies to be acknowledged for preparing this paper.

Reference

Chaudhari, A., Kadam, A., Mane, R., Rajam, N., & Khot, A. (2025). Water quality predictionand suggesting filter using machine learning. Proceedings of the 2025 3rdInternationalConference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT). https://doi.org/10.1109/IDCIOT64235.2025.10914891

Danush, G., Pradeeshwar, A., Tharun Kumar, D. S., Krishna, V. N. K., & Vara Prasad, D. V.(2023). Machine learning and ensemble techniques for water quality analysis of Korattur Lake. Proceedings of the 2023 4th IEEE Global Conference for Advancement in Technology (GCAT). https://doi.org/10.1109/GCAT59970.2023.10353317

David, J., & Shibin, D. (2024). Adaptive water quality potability prediction and analysisthrough GEM continual learning algorithm for sustainable resource management. Proceedings of the 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS). https://doi.org/10.1109/ICACCS60874.2024.10717036

Dumbre, D., Devi, S., & Chavan, R. (2024). Predicting water safety: Harnessing the power ofsimple machine learning algorithms. Proceedings of the 2024 International Conference on Advances in Computing, Communication, and Artificial Intelligence (ACCAI).https://doi.org/10.1109/ACCAI61061.2024.10602429

Gurumurthy, M., & Chitra, K. (2024). Email Phishing Detection Model using CNN Model. Journal of Innovation and Technology, 2024. https://doi.org/10.61453/joit.v2024no43

Han, Y., Pan, Q., & Sun, J. (2024). Prediction of drum water level of power plant boiler basedon random forest. Proceedings of the 2024 3rd International Conference on Energy, Power and Electrical Technology (ICEPET). https://doi.org/10.1109/ICEPET61938.2024.10626447

Huynh, D.-T., & Do, T.-H. (2024). Salinity prediction of raw water using deep learning basedtime series model. Proceedings of the 2024 International Conference on InformationNetworking (ICOIN), 208–213. https://doi.org/10.1109/ICOIN59985.2024.10572127

Meenalochani, M., Hariharan, T., & Kumar, S. V. (2024). Edge-based device using machinelearning for water quality management in a smart campus. Proceedings of the 2024IEEE International Conference on Electronics, Communication and AerospaceTechnology (ICECA). https://doi.org/10.1109/ICECA63461.2024.10800837

Rao, L. G., Sravani, K., Nanre, D., Nethani, S., & Mohmmad, S. (2024). An optimised modelto forecast the water potability using the restricted Boltzmann machine and neural network. Proceedings of the 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT). https://doi.org/10.1109/IC2PCT60090.2024.10486286

Tummala, M., Ajith, K., Mamidibathula, S. K., & Kenchetty, P. (2024). Driving sustainablewater solutions: Optimizing water quality prediction using a stacked hybrid model with gradient boosting and ridge regression. Proceedings of the 2024 3rd InternationalConference on Automation, Computing and Renewable Systems (ICACRS). https://doi.org/10.1109/ICACRS62842.2024.10841518