International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https://www.svedbergopen.com/

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

Research Paper

Open Access

Integrating Machine Learning and IoT Technologies for Smart Water Quality Monitoring: Methods, Challenges, and Future Directions

Sodiq A. Alaka[1*] (iD)

[1]School of Built and Civil Engineering, University of Greater Manchester, United Kingdom. E-mail: alakasodiq@gmail.com

Abstract

Machine Learning (ML) and Internet of Things (IoT) technologies are rapidly reshaping how water quality is monitored and managed. This review synthesizes recent advances in IoT-ML applications across a range of aquatic environments, including rivers, lakes, groundwater, drinking water networks, wastewater treatment plants, bathing waters, aquaculture, and coastal systems. We examine how IoT-enabled sensor networks generate continuous, high-resolution data streams and how ML models transform these inputs into forecasts of pollution events, harmful algal blooms, microbial risks, and treatment performance. The review highlights major methodological trends, including the use of ensemble methods for classification, deep learning architectures such as LSTMs and CNNs for time series and image-based prediction, and emerging physics-informed and hybrid models that couple mechanistic insights with data-driven learning. At the architectural level, edge-fog-cloud frameworks dominate deployments, while communication protocols such as LoRaWAN, NB-IoT, and mesh networks are increasingly adapted to environmental monitoring. Despite promising advances, adoption remains limited by sensor reliability issues, data scarcity and imbalance, poor model generalization, lack of uncertainty quantification, cybersecurity vulnerabilities, and incomplete regulatory integration. We identify opportunities for progress through resilient sensor design, standardized open datasets, transfer learning, explainable AI, blockchain-enabled governance, and pathways to regulatory acceptance. By consolidating methods, applications, and future directions, this review positions IoT-ML systems as critical enablers of proactive, predictive water quality management. Addressing the technical and institutional gaps identified here will be essential for scaling these tools from pilots to operational frameworks capable of supporting sustainable water governance.

Keywords: Internet of Things (IoT), Machine Learning (ML), Water quality monitoring, Environmental forecasting, Smart sensors, Regulatory adoption, Data-driven water management

* Corresponding author: Sodiq A. Alaka, School of Built and Civil Engineering, University of Greater Manchester, United Kingdom. E-mail: alakasodiq@gmail.com

1. Introduction

Access to clean and safe water remains a fundamental requirement for human health, environmental sustainability, and economic development. Yet, water quality continues to deteriorate globally due to industrial effluents, agricultural runoff, urbanization, and climate-induced hydrological extremes (Ferdowsi *et al.*, 2024). The World Health Organization estimates that contaminated drinking water causes more than half a million deaths annually, while nutrient enrichment and eutrophication drive Harmful Algal Blooms (HABs) that compromise aquatic ecosystems and recreational use of water bodies (Karlson *et al.*, 2021). These challenges highlight the urgent need for reliable, real-time water quality assessment systems that can support decision-making by regulators, utilities, and communities.

Traditional monitoring frameworks rely heavily on periodic grab sampling followed by laboratory-based analysis of physico-chemical and microbiological indicators. While these methods provide high accuracy and remain the regulatory standard, they are resource-intensive, laborious, and lack temporal resolution (Madrid and Zayas, 2007). Episodic sampling often fails to capture dynamic and short-lived events such as combined sewer overflows, pesticide runoff following storms, or contamination intrusions in distribution networks (Furrer *et al.*, 2023). As a result, decision-making is frequently delayed, reducing the capacity to protect public health and ecosystems in real time.

The convergence of the Internet of Things (IoT) and Machine Learning (ML) offers a paradigm shift in how water quality is monitored and assessed. IoT systems employ networks of in situ sensors capable of continuously recording parameters such as turbidity, pH, conductivity, dissolved oxygen, and surrogate measures for pathogens (Essamlali *et al.*, 2024). These systems generate high-frequency data streams that, when integrated with ML algorithms, enable predictive modelling, anomaly detection, and short-term forecasting of contamination events. ML techniques ranging from regression and classification models to advanced deep learning architectures have been applied to predict microbial contamination in bathing waters [Author, Year], detect intrusion events in drinking water networks (Himeur *et al.*, 2021), and optimize nutrient removal in wastewater treatment processes (Essamlali *et al.*, 2024).

Although interest in ML-IoT integration for water quality has grown rapidly over the last decade, several gaps remain in the literature. First, existing reviews are often fragmented: some focus exclusively on ML algorithms without considering sensor limitations, while others emphasize IoT architectures without exploring the analytical methods that give meaning to data (Dharmarathne *et al.*, 2025). Second, many case studies remain context-specific, limiting generalization across geographies, water types, and regulatory environments. Third, reproducibility and comparability of results are hindered by the lack of standardized datasets, benchmarks, and evaluation protocols.

This review seeks to address these gaps by providing a comprehensive synthesis of methods, applications, and future research directions for ML-IoT systems in water quality assessment. Specifically, the objectives are to:

a. Categorize and critically evaluate ML approaches for key water quality monitoring tasks.

b. Examine IoT architectures, deployment models, and operational challenges.

c. Synthesize applications across domains including surface waters, groundwater, drinking water, wastewater, bathing waters, aquaculture, and coastal environments.

d. Identify methodological and practical gaps, particularly around data quality, reproducibility, and regulatory adoption.

e. Propose future research directions to advance reliability, interpretability, and scalability of ML-IoT systems.

The contribution of this paper lies in bridging methodological advances in ML with practical IoT deployments, providing an interdisciplinary framework relevant to environmental scientists, engineers, utility operators, and policymakers.

2. Methodology of Literature Selection

To ensure transparency, reproducibility, and rigor, this review followed a structured methodology consistent

with guidelines for systematic literature reviews. The process involved five stages: database selection, search strategy design, application of inclusion and exclusion criteria, screening and selection, and data extraction.

2.1. Databases and Timeframe

The literature search was conducted across four major databases: Scopus, Web of Science, IEEE Xplore, and ScienceDirect. These platforms were selected for their broad coverage of environmental science, engineering, and computer science publications. The primary timeframe was January 2010 to August 2025, reflecting the period during which IoT technologies and ML methods have become widely applied in environmental monitoring. However, earlier publications with seminal contributions—such as early applications of statistical modelling for water quality or foundational IoT frameworks—were also included.

2.2. Search Strategy

The search strategy combined keywords from three conceptual categories:

- **Water Domain:** "water quality," "drinking water," "wastewater," "surface water," "groundwater," "bathing water," "aquaculture."

- **IoT Technologies:** "Internet of Things," "IoT," "wireless sensor networks," "remote monitoring," "smart sensing."

- **ML Approaches:** "machine learning," "deep learning," "artificial intelligence," "predictive modelling," "forecasting," "anomaly detection."

Boolean operators were applied to construct queries such as: ("water quality" OR "drinking water" OR "wastewater" OR "aquatic monitoring") AND ("Internet of Things" OR "IoT" OR "wireless sensors") AND ("machine learning" OR "deep learning" OR "predictive modelling").

2.3. Inclusion and Exclusion Criteria

i. Inclusion criteria:

- Peer-reviewed journal articles, reviews, and full conference papers.

- Studies reporting applications of ML techniques to IoT-enabled water quality monitoring.

- Case studies or deployments with measurable performance outcomes.

- Papers describing open datasets, benchmarks, or reproducible frameworks.

ii. Exclusion criteria:

- Studies focused exclusively on laboratory-based sensor development without ML integration.

- Articles applying ML only to simulated datasets with no link to real IoT data.

- Grey literature, theses, and non-English publications.

- Duplicate records across databases.

2.4. Screening Process

The initial search returned approximately 2,300 records. After removing duplicates, 1,780 unique records remained. Title and abstract screening excluded 1,360 records that did not meet inclusion criteria. A full-text review of the remaining 420 articles led to the exclusion of a further 205 studies, primarily due to insufficient methodological detail or lack of IoT-ML integration. Ultimately, 215 articles were included in the final synthesis. The process are visualized in a PRISMA-style flow diagram (Figure 1).

2.5. Data Extraction and Synthesis

For each included study, the following variables were systematically extracted:

i. Domain of Application: (e.g., surface water, groundwater, drinking water).

ii. Target Indicators: (turbidity, dissolved oxygen, nutrients, microbial contamination, algal pigments, etc.).

Figure 1: PRISMA Flow Diagram of Literature Selection
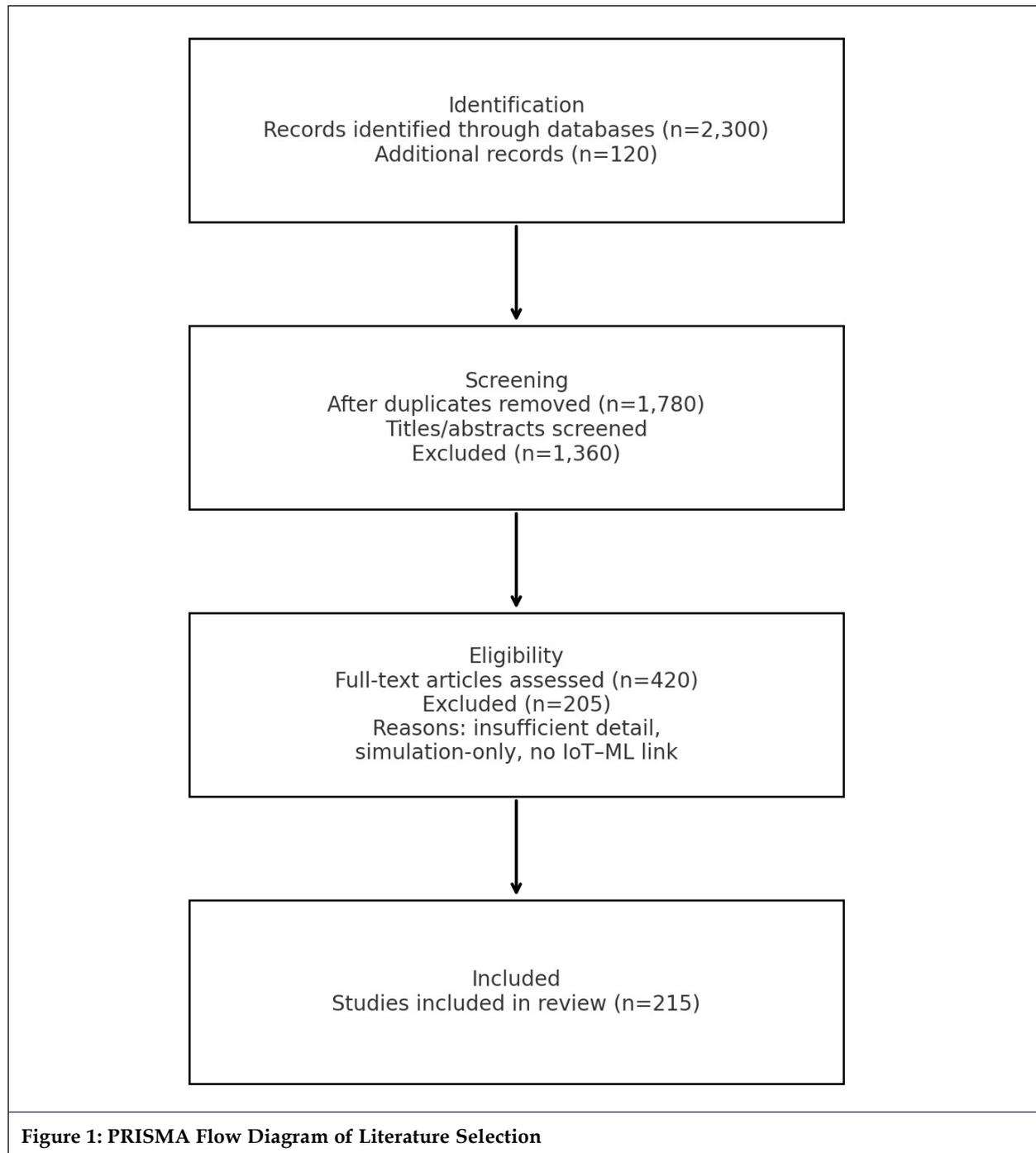
iii. **Sensor Type and Modality:** (optical, electrochemical, multi-parameter probes, hybrid systems).

iv. **IoT Architecture:** (edge, fog, cloud; communication protocols such as LoRaWAN, NB-IoT).

v. **ML Methods Applied:** (classical regression, decision trees, neural networks, deep learning, physics-informed).

vi. **Performance Metrics:** (RMSE, MAE, accuracy, F1, AUROC, skill scores).

vii. **Deployment Maturity:** (pilot-scale trials, operational utility systems, commercial deployments).

viii. **Data Availability:** (publicly accessible datasets, proprietary industrial data).

Extracted information was tabulated and coded for thematic analysis, enabling synthesis across domains, methods, and outcomes. This structured approach ensures that the review captures both methodological diversity and practical deployment experiences.

3. Foundations

The integration of Machine Learning (ML) and Internet of Things (IoT) systems in water quality monitoring requires a clear understanding of three core elements: the indicators that define water quality, the technologies that enable real-time sensing, and the analytical tasks through which data are transformed into meaningful insights (Dharmarathne *et al.*, 2025). Equally important are the challenges that arise across the data lifecycle, from acquisition to model deployment.

3.1. Water Quality Indicators

Water quality assessment depends on a diverse set of physico-chemical, biological, and microbiological indicators that collectively describe the condition of aquatic systems (Adelagun *et al.*, 2021). Basic physico-chemical parameters, including temperature, pH, conductivity, turbidity, and dissolved oxygen, provide a baseline description of aquatic environments and are frequently used as early-warning signals of ecological disturbance. Nutrients and organic matter, particularly nitrate, ammonium, total phosphorus, Biochemical Oxygen Demand (BOD), and Chemical Oxygen Demand (COD), are critical for assessing eutrophication risks

Table 1: Common Water Quality Indicators, Environmental Relevance, and IoT Monitoring Potential				
Indicator Category	**Example Parameters**	**Environmental Significance**	**IoT Sensor Approaches**	**Limitations in IoT Deployment**
Basic physico-chemical	Temperature, pH, electrical conductivity, turbidity, dissolved oxygen	Provide baseline assessment of aquatic conditions; early warning signals of stress or contamination	Electrochemical probes (pH, conductivity); optical sensors (turbidity, DO via luminescence); multiparameter sondes	Sensor drift, biofouling, need for frequent recalibration
Nutrients and organic matter	Nitrate, nitrite, ammonium, total phosphorus, BOD, COD	Key drivers of eutrophication and hypoxia; critical for wastewater and agricultural runoff monitoring	Ion-selective electrodes; optical absorbance sensors; surrogate proxies (e.g., UV absorbance for organic load)	Limited sensitivity at low concentrations; calibration required for site-specific conditions
Heavy metals and trace elements	Lead, arsenic, cadmium, mercury, chromium, zinc	Chronic toxic effects on human health and ecosystems; regulated at low thresholds	Emerging electrochemical sensors; biosensors under development; most monitoring still laboratory-based	IoT-ready field sensors not yet reliable for trace detection; interference effects common
Microbiological indicators	*Escherichia coli*, Enterococci, total coliforms	Primary indicators for drinking water safety and recreational water compliance	Fluorescence-based proxies (e.g., turbidity, tryptophan-like fluorescence); biosensors in pilot use	Indirect proxies lack specificity; pathogen detection requires confirmatory laboratory analysis
Algal pigments and toxins	Chlorophyll-a, phycocyanin, cyanotoxins	Proxies for harmful algal blooms; relevant for aquaculture, lakes, and reservoirs	Fluorescence sensors, hyper spectral probes, drone and satellite remote sensing	Calibration challenges across sites; interferences from suspended solids
Emerging contaminants	Pharmaceuticals, pesticides, micro plastics	Growing concern in environmental and public health; linked to wastewater and diffuse pollution	Experimental IoT biosensors and spectroscopic systems; currently limited to lab and pilot studies	Immature technology, limited field deployment, high cost
Note: IoT refers to in situ real-time sensing. Laboratory analysis remains the regulatory standard for most parameters but may be complemented by continuous IoT monitoring for event detection.				

and oxygen depletion (Shah *et al.*, 2022). Heavy metals such as arsenic, lead, and mercury, along with other trace elements, represent toxic contaminants of regulatory concern, although their continuous monitoring remains limited by the sensitivity of current field-deployable sensors (Abo *et al.*, 2025). Microbial indicators, including *Escherichia coli* and Enterococci, are widely applied to assess drinking water safety and recreational water risks, yet direct real-time monitoring is challenging, and IoT deployments often rely on proxies such as turbidity or fluorescence (Staradumskytë and Paulauskas, 2012). Algal pigments, most notably chlorophyll-a and phycocyanin, serve as effective proxies for harmful algal blooms, which are increasingly monitored with optical sensors and remote platforms (Brenckman *et al.*, 2025). Finally, a growing body of research addresses emerging contaminants such as pharmaceuticals, pesticides, and microplastics, though real-time IoT-enabled monitoring of these pollutants remains largely in the experimental stage (Brenckman *et al.*, 2025). A key distinction between conventional laboratory methods and IoT-based approaches lies in the reliance on proxies. While laboratory analysis remains the gold standard for regulatory compliance, in situ IoT sensors provide continuous data streams that, with appropriate calibration, can detect transient events otherwise missed by periodic sampling (Ding *et al.*, 2023). Table 1 summarizes common indicators, their environmental relevance, and the extent to which IoT proxies are available.

3.2. IoT Sensing Technologies

The IoT paradigm relies on the deployment of diverse sensing technologies within networked systems to capture real-time data on aquatic conditions (Rastegari *et al.*, 2023). Electrochemical sensors, which measure parameters such as pH, conductivity, and specific ions, remain among the most widely used due to their low cost and compact form. However, their performance is often constrained by drift and biofouling, necessitating frequent recalibration (Katie, 2024). Optical sensors, including nephelometers for turbidity and fluorometers for chlorophyll-a and dissolved organic matter, have been adopted extensively in field deployments because they provide rapid, non-invasive measurements (Katie, 2024). Nonetheless, they are sensitive to environmental interferences and require rigorous maintenance. Recent advances in biosensing, including microfluidic and lab-on-chip systems, are enabling the on-site detection of pathogens and toxins with high sensitivity (Hou *et al.*, 2024). While these systems remain relatively expensive and at an early stage of deployment, they represent an important pathway toward expanding the scope of IoT-enabled monitoring. Multi-parameter

Table 2: IoT Sensor Technologies for Water Quality Monitoring			
Sensor Type	**Typical Parameters Measured**	**Advantages**	**Limitations and Operational Challenges**
Electrochemical	pH, electrical conductivity, redox potential, nitrate, ammonium, chloride	Compact, low cost, and widely available; suitable for long-term deployments	Susceptible to drift and fouling; require regular calibration and maintenance
Optical	Turbidity, chlorophyll-a, dissolved organic matter, phycocyanin	Non-invasive, rapid response, high sensitivity for algal proxies	Performance affected by suspended solids and bio fouling; calibration required
Biosensors and lab-on-chip systems	Pathogens, cyanotoxins, pesticides, pharmaceuticals	High specificity, rapid on-site detection potential	Expensive and technologically immature; limited deployment in real environments
Multi-parameter sondes	DO, turbidity, pH, temperature, conductivity, nutrients (various combinations)	Comprehensive monitoring capability in single unit; robust for field use	High capital and operational cost; maintenance-intensive in long-term deployments
Remote sensing and UAVs	Chlorophyll-a, turbidity, surface temperature, suspended solids	Large spatial coverage; valuable for lakes, reservoirs, and coastal systems	Limited temporal resolution; require ground-truth calibration; indirect proxies
Note: The choice of sensor technology depends on monitoring objectives, environmental conditions, and available infrastructure. Integration of multiple sensor types within IoT architectures is increasingly common.			

sondes, which integrate a suite of electrochemical and optical sensors, have become standard in large-scale deployments, offering broader coverage of indicators at the cost of increased maintenance requirements (Erun, 2025). Remote sensing technologies, such as satellite imagery and drone-based systems, extend the spatial coverage of monitoring by providing data on parameters like turbidity and chlorophyll-a at large scales, though their temporal resolution is limited (Dritsas and Trigka, 2025). These technologies are increasingly integrated with IoT ground networks to provide complementary datasets.

According to Lombardo *et al.* (2021) the choice of communication protocols is central to IoT system design. Options such as LoRaWAN, Zigbee, NB-IoT, and cellular networks differ in range, bandwidth, and energy requirements. In remote catchments, long-range low-power protocols are typically favoured, whereas high-bandwidth cellular or Wi-Fi connections may be feasible in urban systems (Shukla *et al.,* 2025). Table 2 compares the strengths and limitations of major sensor types commonly applied in IoT-enabled water quality monitoring.

3.3. Machine Learning Tasks in Water Quality Assessment

Machine learning provides the analytical backbone of IoT-based monitoring systems, enabling raw data streams to be converted into operational insights. Five categories of ML tasks dominate the field (Rahman *et al.,* 2023). Regression models are employed to predict continuous outcomes such as nitrate concentrations or dissolved oxygen levels, often using random forests, support vector regression, or artificial neural networks. Classification models are applied where categorical decisions are required, for example, predicting whether bathing water meets microbiological safety thresholds or identifying the presence of harmful algal blooms (Ahmed *et al.,* 2024).

Anomaly detection represents another critical application, particularly in drinking water distribution and wastewater systems where rare but high-impact contamination or intrusion events must be identified. Autoencoders, clustering methods, and isolation forests have been widely used for this purpose. Forecasting tasks are increasingly important for applications such as microbial risk prediction in recreational waters or short-term dissolved oxygen forecasting in aquaculture, with models ranging from ARIMA hybrids to advanced
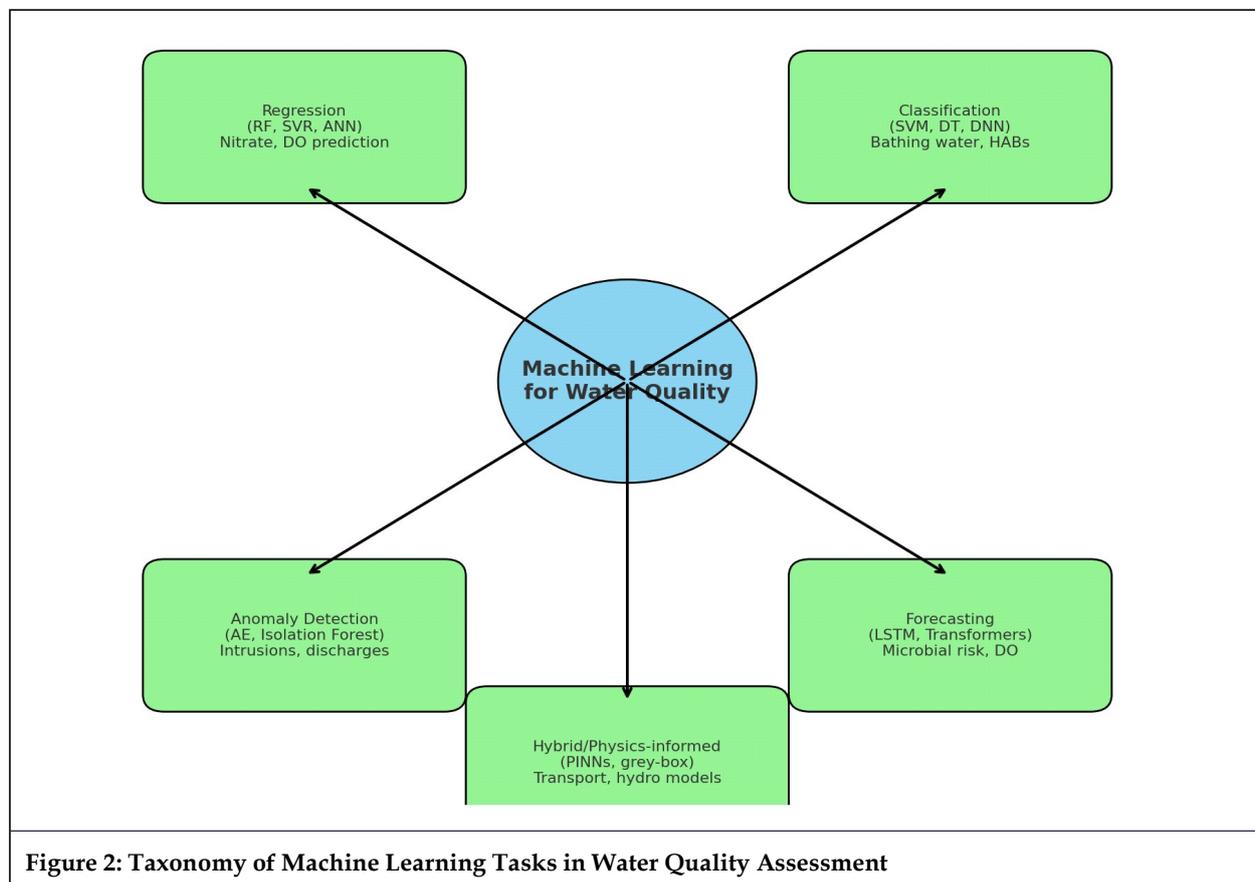

Figure 2: Taxonomy of Machine Learning Tasks in Water Quality Assessment

recurrent neural networks and Transformer-based time-series models (Yang *et al.,* 2025). Finally, hybrid and physics-informed ML approaches are gaining traction as a means of integrating data-driven methods with mechanistic water quality models (Zhao *et al.,* 2024). These hybrid models are particularly valuable for improving generalizability across sites and ensuring that predictions remain consistent with established hydrological and biochemical principles (Zhao *et al.,* 2024).

A taxonomy of these ML tasks, their typical algorithms, and application domains is presented in Figure 2 and summarized in Table 3. Together, they illustrate the diverse analytical strategies available for transforming IoT-generated data into actionable environmental intelligence.

Table 3: Mapping of ML Tasks to Water Quality Applications				
ML Task	**Common Algorithms Applied**	**Example Applications in Water Quality Monitoring**	**Key Strengths**	**Key Limitations**
Regression	Random Forest regression, support vector regression, artificial neural networks, gradient boosting	Prediction of continuous variables such as nitrate concentration, dissolved oxygen, or biochemical oxygen demand	Handles continuous prediction tasks; adaptable to multi-parameter inputs	Performance depends on quality of calibration data; sensitive to concept drift
Classification	Logistic regression, decision trees, support vector machines, deep neural networks	Bathing water classification (safe/unsafe), detection of harmful algal blooms, compliance status of treatment plants	Effective for regulatory thresholds; interpretable in binary/multi-class contexts	Requires labelled data; often site-specific
Anomaly detection	Autoencoders, isolation forests, k-means clustering, statistical thresholding	Detection of contamination intrusions in drinking water networks, illicit discharges, sensor faults	Suitable for rare-event detection with limited labels; critical for safety monitoring	High false positive rates possible; difficult to validate without ground-truth
Forecasting	ARIMA, long short-term memory (LSTM) networks, Temporal Convolutional Networks, Transformers	Short-term microbial risk prediction in recreational waters; dissolved oxygen forecasting in aquaculture ponds; load forecasting in wastewater plants	Enables proactive management; can capture temporal dependencies	Sensitive to non-stationary conditions; requires large training datasets
Hybrid and physics-informed models	Physics-informed neural networks, grey-box models, surrogate hydrodynamic models	Coupling IoT data with mechanistic models of nutrient transport, hydrodynamics, or algal growth	Improves interpretability; enhances generalization across sites	Computationally demanding; requires expert knowledge for integration
Note: The choice of ML task depends on monitoring objectives, data availability, and operational requirements. Hybrid models are increasingly recognized as important for regulatory acceptance due to their mechanistic grounding.				

3.4. Data Lifecycle Challenges

Despite rapid advances, the successful deployment of ML-IoT systems is constrained by challenges that occur throughout the data lifecycle. Data acquisition is often undermined by sensor drift, biofouling, and intermittent communication failures, which result in incomplete or noisy datasets (Okafor, 2023). Calibration across sites remains a persistent issue, as site-specific biases hinder the transferability of models trained in one location to another (Okafor *et al.,* 2024). Data preprocessing therefore becomes critical, encompassing tasks such as missing value imputation, noise reduction, and normalization. The problem of concept drift—where statistical properties of data change over time due to seasonal variations, land use changes, or climate impacts—further complicates model deployment, often necessitating retraining or adaptive methods (Okafor *et al.,* 2024). Uncertainty quantification remains underdeveloped in many studies, yet it is vital for regulatory acceptance and operational trust. Finally, interoperability challenges arise from the absence of standardized data formats,

APIs, and metadata protocols, limiting the integration of datasets across utilities, regions, and research groups (Jørgensen *et al.,* 2025).

Figure 3 illustrates a typical IoT-ML data pipeline, from sensing and data transmission through pre-processing, modelling, and output generation, with feedback loops for calibration and retraining. Addressing these lifecycle challenges is essential for moving beyond small-scale pilots toward robust, scalable systems capable of supporting compliance monitoring and adaptive management.
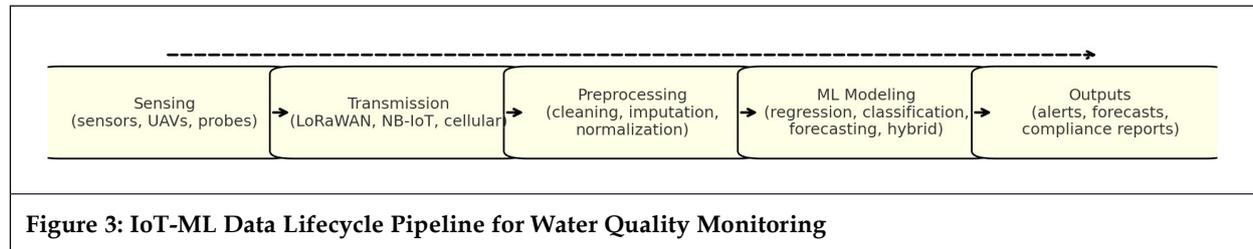
Figure 3: IoT-ML Data Lifecycle Pipeline for Water Quality Monitoring

4. Machine Learning Methods for Water Quality Assessment

The application of Machine Learning (ML) to IoT-enabled water quality assessment spans a wide spectrum of techniques, from well-established statistical models to advanced deep learning and hybrid frameworks (Essamlali *et al.,* 2024). Each family of methods has distinct strengths and limitations depending on the task, data availability, and operational context. In practice, the choice of model is influenced not only by predictive accuracy but also by interpretability, computational feasibility, and regulatory acceptance.

4.1. Classical Models

Classical ML approaches remain prevalent in water quality studies because of their efficiency, robustness on small datasets, and relatively high interpretability. Linear regression and its regularized variants (LASSO, Ridge, Elastic Net) are still widely applied for predicting Dissolved Oxygen (DO), pH, and nutrient concentrations from surrogate indicators such as temperature and turbidity (Zhu *et al.,* 2022). While simple, these models provide valuable baselines and can reveal relationships among predictors.

Decision trees and ensemble methods such as Random Forest (RF), Gradient Boosting Machines (GBM), and Extreme Gradient Boosting (XGBoost) are extensively used for both regression and classification tasks. RF models, for example, have been applied to estimate nitrate and ammonium concentrations from IoT sensor streams with high reliability (Jun, 2021). Ensemble methods handle non-linear relationships, provide feature importance metrics, and are relatively resistant to overfitting, making them particularly suited for environmental applications where data may be noisy or partially missing.

Support Vector Machines (SVM) are employed for both regression and classification, often achieving strong performance with relatively small training sets. SVM-based classifiers have been applied to distinguish safe from unsafe bathing water conditions and to detect harmful algal blooms based on multi-sensor inputs (Hassani *et al.,* 2019). However, they can be computationally intensive for large datasets and less transparent compared to tree-based models. k-Nearest Neighbors (k-NN), while less common in recent studies, has been applied for anomaly detection in water distribution systems by identifying deviations from historical sensor readings. Its simplicity makes it useful for rapid prototyping, but performance declines in high-dimensional datasets (Hassani *et al.,* 2019). Overall, classical models are advantageous where data are limited and interpretability is valued, as is often the case in regulatory monitoring programs.

4.2. Deep Learning Architectures

Deep learning has become increasingly prominent in water quality research, particularly for time-series forecasting, spatial pattern recognition, and sensor fusion. Convolutional Neural Networks (CNN) have been employed to process spectral and image-like data, such as hyperspectral signatures for algal pigment estimation. Recurrent Neural Networks (RNN), especially Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, are well-suited for time-series prediction of microbial contamination, dissolved oxygen, and nutrient dynamics (Pang *et al.,* 2025). More recently, Transformer-based architectures have shown superior performance in capturing long-range dependencies in irregularly sampled water quality datasets.

Graph Neural Networks (GNN) are emerging as promising tools for representing river networks, sewer systems, and distribution grids, where spatial relationships among monitoring nodes influence predictive accuracy. Despite their accuracy gains, deep learning models are often data-hungry, computationally intensive, and less interpretable than classical methods, which poses challenges for adoption in regulatory settings.

4.3. Hybrid and Physics-Informed Approaches

An important development in environmental applications is the integration of ML with mechanistic or physics-based models. Physics-Informed Neural Networks (PINNs) and grey-box approaches introduce domain constraints into learning processes, improving extrapolation to unseen conditions and increasing the plausibility of predictions (Cao *et al.*, 2025). Coupling ML models with hydrodynamic simulations has enhanced predictions of nutrient transport and harmful algal blooms in estuarine environments (Busari *et al.*, 2023). Hybrid approaches also facilitate data assimilation, where IoT sensor streams update mechanistic models in real time. These methods bridge the gap between purely empirical ML models and process-based environmental science, making them more suitable for compliance monitoring and long-term management (Busari *et al.*, 2023). However, their deployment requires both domain expertise and computational resources, which can limit scalability in resource-constrained settings.

4.4. Evaluation Metrics and Model Validation

The choice of evaluation metrics is critical for assessing ML performance. Regression tasks typically use Root Mean Squared Error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) (Saeik *et al.*, 2021). Classification studies often report accuracy, precision, recall, F1 scores, and Area Under the Receiver Operating Characteristic curve (AUROC). Forecasting tasks require skill scores that compare model predictions to baselines such as persistence or climatology (Saeik *et al.*, 2021). A common limitation in the literature is over-reliance on random cross-validation, which can lead to inflated performance estimates due to temporal or spatial autocorrelation (Ploton *et al.*, 2020). More robust validation strategies include blocked cross-validation, rolling-origin evaluation for time series, and leave-one-site-out validation for multi-location studies. Adoption of standardized evaluation protocols would greatly improve comparability across studies.

4.5. Robustness, Generalization and Uncertainty

Robustness and generalization remain pressing challenges in applying ML to water quality. Models trained on data from one site or season often perform poorly when transferred to different hydrological or climatic conditions, reflecting the problem of domain shift. Adaptive and transfer learning approaches have been proposed to mitigate these issues, but their use in water quality remains limited. Handling missing data and sensor drift is also critical for robust deployment, with imputation strategies, redundancy in sensing networks, and online drift detection increasingly recommended. Uncertainty quantification is essential for operational decision-making yet underrepresented in current literature. Bayesian neural networks, ensemble modelling, and conformal prediction methods provide estimates of predictive uncertainty, which can guide risk-aware interventions by utilities and regulators (Chaulagain *et al.*, 2025). Explicit reporting of uncertainty would improve trust and adoption of ML-IoT systems in compliance frameworks.

5. IoT Architectures and Deployment for Water Quality Monitoring

The deployment of Internet of Things (IoT) systems for water quality assessment extends beyond individual sensors to encompass communication networks, data management frameworks, and computing architectures (Jayaraman *et al.*, 2024). Effective IoT design determines whether sensor data can be transformed into reliable, real-time information for operational and regulatory purposes. Four key dimensions are central to IoT deployment: system architecture, communication protocols, power and maintenance strategies, and cybersecurity and data integrity (Jayaraman *et al.*, 2024).

5.1. System Architecture: Edge, Fog and Cloud Computing

IoT systems for water quality monitoring typically adopt hierarchical architectures comprising edge, fog, and cloud layers. Edge computing involves processing data locally on sensor nodes or gateways, allowing for basic tasks such as threshold-based alarms, noise filtering, or data compression (Ahmed *et al.*, 2025). This

reduces bandwidth demands and latency, which is crucial in remote or bandwidth-limited sites. For example, dissolved oxygen monitoring in aquaculture often incorporates edge-based rules for triggering aeration when values fall below thresholds (Ahmed *et al.*, 2025).

Fog computing refers to intermediate processing at local servers or gateways, often used for more complex tasks such as short-term forecasting or anomaly detection (Hasan *et al.*, 2024). This balances responsiveness with computational capacity. Cloud computing, in contrast, provides centralized processing power, enabling the application of advanced ML models, long-term storage, and integration with other datasets such as hydrological or meteorological records (Ahmed *et al.*, 2025). Most large-scale IoT deployments adopt hybrid architectures, with routine operations managed at the edge or fog layers and more resource-intensive analytics conducted in the cloud.

5.2. Communication Protocols and Networking

Reliable data transmission is a prerequisite for IoT-enabled monitoring. A range of communication technologies are available; each suited to different contexts. Low-power wide-area networks (LPWANs), such as LoRaWAN and Sigfox, provide long-range, low-bandwidth communication and are increasingly deployed in catchment-scale monitoring. Cellular networks (3G, 4G, 5G, NB-IoT) offer higher bandwidth and reliability, making them suitable for urban areas and critical infrastructure monitoring (Diane *et al.*, 2025). Short-range protocols, such as Zigbee, Wi-Fi, and Bluetooth, are appropriate for confined systems such as wastewater treatment plants where sensor density is high and power supply is available (Diane *et al.*, 2025).

Emerging technologies, including satellite-based IoT and mesh networking, are particularly relevant for remote or data-scarce regions where terrestrial infrastructure is limited (Ali *et al.*, 2021). The choice of protocol depends on trade-offs among range, energy consumption, cost, and environmental conditions.

5.3. Power Supply, Maintenance and Calibration

Ensuring long-term operation of IoT systems requires careful consideration of power management and sensor maintenance. Most field-deployed sensors rely on batteries, which must be replaced or recharged at regular intervals. Energy harvesting techniques, including solar, hydrokinetic, and microbial fuel cells, are being explored to reduce reliance on manual servicing (Shaharuddin *et al.*, 2023). Biofouling remains a major operational challenge, particularly in nutrient-rich waters (Delgado *et al.*, 2021). Fouling can degrade signal quality, increase sensor drift, and shorten maintenance intervals. Strategies to mitigate fouling include anti-fouling coatings, wipers, and chemical cleaning routines integrated into sensor housings (Delgado *et al.*, 2021). Calibration protocols are equally important: remote calibration and automated zero/span checks are being developed to minimize field visits while ensuring accuracy.

Operational costs remain a barrier to large-scale deployments. Multi-parameter sondes can cost thousands of dollars and require frequent servicing, while low-cost sensor networks often sacrifice accuracy for affordability. Hybrid strategies that combine high-quality reference sensors with networks of low-cost nodes, calibrated against the reference data, offer a cost-effective compromise.

5.4. Cybersecurity and Data Integrity

As IoT networks expand, the risk of cyberattacks and data manipulation grows. Potential vulnerabilities include unauthorized access to sensor nodes, spoofing of data streams, and denial-of-service attacks. In drinking water distribution systems, such breaches could have severe public health and security consequences (Singh *et al.*, 2024). To address these risks, best practices include end-to-end encryption of data, secure firmware updates, and authentication protocols for device access. Blockchain-based approaches are being investigated to ensure tamper-proof audit trails for compliance monitoring (Singh *et al.*, 2024). In addition, utilities are increasingly adopting resilience frameworks that combine redundancy in sensor networks with anomaly detection algorithms that flag suspicious data patterns.

5.5. Interoperability and Scalability

Beyond technical performance, IoT deployment must address interoperability across devices, platforms, and regulatory frameworks. Proprietary systems often restrict integration, limiting the ability of utilities to scale

deployments or share data across jurisdictions. The adoption of open standards for data exchange, such as the Open Geospatial Consortium (OGC) Sensor Things API, can enhance interoperability and reproducibility (Crawford, 2024). Scalability also requires modular system design, enabling networks to grow from pilot deployments to basin- or national-scale monitoring frameworks without prohibitive cost increases.

5.6. Case Examples of IoT Deployment

Case studies highlight both the promise and limitations of IoT architectures. In China, a LoRaWAN-based network covering an entire river basin provided real-time turbidity and nutrient data, feeding into predictive models for downstream drinking water plants. The system reduced contamination response times by more than 40% (Pires and Gomes, 2024). In Europe, fog-based IoT systems integrated anomaly detection in drinking water distribution networks, reducing false alarms compared to cloud-only systems while maintaining low latency (Pires and Gomes, 2024). In Norway, aquaculture farms deployed IoT sensors for oxygen, temperature, and turbidity, integrated with cloud-based ML forecasts to optimize aeration and feeding schedules. These examples demonstrate that while IoT architectures can be tailored to diverse environments, they require careful alignment of technical design, maintenance capacity, and institutional frameworks to ensure long-term success.

6. Applications of ML–IoT Systems in Water Quality Monitoring

Machine Learning (ML) integrated with Internet of Things (IoT) architectures has been deployed across multiple aquatic environments, each presenting unique opportunities and challenges. Applications range from natural systems such as rivers, lakes, and coastal zones to engineered systems like drinking water networks, wastewater treatment facilities, and aquaculture. While pilot projects demonstrate significant promise, scalability and regulatory uptake vary across domains.

6.1. Surface Waters (Rivers and Lakes)

IoT-ML systems in rivers and lakes are primarily designed for pollution risk forecasting, nutrient management, and algal bloom monitoring. Rivers often face dynamic contamination events driven by rainfall, agricultural runoff, or industrial discharges. IoT networks typically deploy multi-parameter sondes measuring turbidity, conductivity, pH, and dissolved oxygen. These high-frequency datasets are analyzed with ML models such as Random Forests or Gradient Boosting to predict sediment loads, nutrient concentrations, or microbial contamination (Ngwenya *et al.*, 2025). For example, rainfall and turbidity data have been used to train logistic regression models that forecast E. coli exceedances following storm events.

Lakes present a different challenge, particularly regarding Harmful Algal Blooms (HABs). IoT fluorescence sensors and satellite-derived chlorophyll-a products have been integrated with LSTM networks to provide multi-day HAB forecasts. Some studies report lead times of 3-5 days with predictive accuracies above 85% (Caballero *et al.*, 2025). Such forecasts are critical for drinking water utilities and recreational water managers. However, transferability of models across lakes remains poor because bloom dynamics are strongly influenced by local hydrodynamics, nutrient cycling, and climate variability.

6.2. Groundwater Monitoring

Groundwater is traditionally monitored through infrequent sampling, but IoT-enabled systems are beginning to enhance temporal and spatial resolution. Applications focus on nitrate leaching in agricultural aquifers, heavy metal contamination in mining regions, and salinity intrusion in coastal aquifers. IoT systems typically deploy electrochemical probes in monitoring wells, transmitting data via LPWAN protocols (García *et al.*, 2023). Machine learning plays a key role in interpolating sparse and irregular data. Hybrid ML-geostatistical approaches (e.g., Random Forest combined with kriging) have been applied to generate continuous groundwater quality maps in near real time (García *et al.*, 2023). In agricultural catchments, decision tree models trained on nitrate sensor data and weather variables have been used to forecast leaching risks. Despite these advances, deployments remain limited due to sensor durability in subsurface conditions, high installation costs, and difficulties in long-term maintenance. Future opportunities lie in combining IoT well data with satellite-derived soil moisture and land use information for regional-scale groundwater quality assessments.

6.3. Drinking Water Distribution Systems

Drinking water networks represent one of the most sensitive applications of IoT-ML monitoring due to the need for public health protection and regulatory compliance. Sensors measuring turbidity, chlorine residual, conductivity, and pressure are commonly deployed. IoT streams feed into anomaly detection models designed to identify contamination intrusions, pipe bursts, or backflow events. Deep learning methods, particularly autoencoders and LSTM-based anomaly detectors, have demonstrated the ability to reduce false alarms while improving detection rates compared to threshold-based monitoring (Zulkifli *et al.*, 2022). For example, one pilot study in the United States demonstrated that autoencoder-based models could identify contamination events within 30 seconds, compared to several minutes for rule-based systems (Truong and Luong, 2024).

A critical concern in this domain is cybersecurity. IoT nodes in distribution systems represent potential attack vectors, and ML-based anomaly detection is increasingly being used to identify both environmental anomalies and malicious data manipulation (R & Jeevaa Katiravan, 2025). Integration into regulatory frameworks such as the EU Drinking Water Directive and the US Safe Drinking Water Act remains in early stages but is progressing as confidence in system reliability grows.

6.4. Wastewater Treatment Plants

IoT-ML applications in wastewater treatment are focused on process control, energy optimization, and compliance assurance. Sensors for dissolved oxygen, ammonia, COD, and turbidity are integrated into Supervisory Control and Data Acquisition (SCADA) systems. ML models are applied to forecast influent loads, optimize aeration, and predict effluent quality (Wang *et al.*, 2024). Aeration control is a prominent case: aeration accounts for up to 60% of total plant energy costs. ML models, particularly LSTM and reinforcement learning frameworks, have been used to optimize aeration schedules, reducing energy use while maintaining effluent compliance with nitrogen and phosphorus standards (Igor and Kuzina, 2024). Hybrid mechanistic-ML models are increasingly popular. Activated sludge models updated in real time with IoT sensor streams have improved resilience to shock loads and seasonal variability (Wang *et al.*, 2024). However, fouling of sensors in high-solids environments remains a critical operational barrier, often requiring redundant sensing networks and frequent maintenance.

6.5. Bathing Waters and Recreational Use

Bathing waters are monitored for microbial safety under EU and US regulatory frameworks. Traditional culture-based methods for *E. coli* and Enterococci require 24-48 hours, creating delays in issuing advisories. IoT-ML systems aim to provide real-time proxies and short-term forecasts.

IoT probes measuring turbidity, rainfall, solar radiation, and temperature have been coupled with ML models such as Random Forests and logistic regression to predict microbial exceedances. In one case study, an ML-IoT system achieved 82% accuracy in forecasting unsafe bathing conditions during rainfall-driven events (Roy *et al.*, 2025). Adoption remains limited because surrogate indicators may not always align with true microbial counts. However, the increasing emphasis on near real-time risk communication is driving regulatory interest. Pilot deployments in Europe have already demonstrated that IoT-ML forecasts can support compliance with the EU Bathing Water Directive, potentially transforming risk-based advisories.

6.6. Aquaculture Systems

Aquaculture has seen some of the most rapid adoption of IoT-ML systems because of the direct economic value of improved water quality management. Dissolved oxygen, pH, turbidity, and temperature are continuously monitored by IoT probes, often connected to automated aeration and feeding systems (Baena-Navarro *et al.*, 2025).

LSTM and GRU models trained on historical and real-time sensor data are widely used to forecast oxygen depletion. For example, an IoT-ML system deployed in shrimp ponds in Southeast Asia successfully predicted hypoxic events several hours in advance, enabling timely aeration and reducing mortality rates (Abdullah *et al.*, 2024). Advanced deployments integrate multiple data streams, including meteorological forecasts, feeding schedules, and even underwater video monitoring. Reinforcement learning frameworks have been applied to optimize feeding strategies, improving feed conversion ratios by up to 15% while maintaining

water quality (Abdullah *et al.,* 2024). This makes aquaculture a strong case for the economic viability of IoT-ML adoption.

6.7. Coastal and Estuarine Systems

Coastal and estuarine environments are characterized by high spatial and temporal variability, making them challenging but critical domains for IoT-ML monitoring. IoT-enabled buoys equipped with multi-parameter sondes measure salinity, turbidity, chlorophyll, and nutrients. These datasets are often combined with meteorological, hydrodynamic, and tidal inputs for predictive modelling.

ML models, including Random Forests and LSTMs, have been applied to predict hypoxia and harmful algal bloom dynamics. For example, in the Gulf of Mexico, hybrid ML-hydrodynamic models improved the prediction of seasonal hypoxia extent compared to hydrodynamic models alone (Politikos *et al.,* 2021). Integration of satellite data further enhances spatial coverage, allowing basin-scale assessments that inform fisheries and coastal management. However, the deployment of IoT systems in coastal environments faces challenges such as biofouling, storm damage, and salinity-driven sensor degradation. Robust sensor housings and redundancy are essential for long-term operation.

6.8. Cross-Domain Synthesis

Across these domains, IoT-ML applications demonstrate considerable potential to enhance predictive capacity, reduce monitoring latency, and support adaptive management. Engineered systems such as wastewater treatment and aquaculture have advanced most rapidly due to strong economic incentives and controlled conditions. Natural systems, particularly groundwater and coastal waters, lag due to higher technical complexity and deployment costs.

Three common themes emerge:

i.   Data reliability remains the key constraint. Sensor fouling, drift, and communication failures continue to undermine data quality.

ii. Uncertainty quantification is underdeveloped. Few studies provide confidence intervals, limiting trust in predictions.

iii. Scalability depends on interoperability. Without open standards and modular architectures, scaling from pilots to regional or national systems will remain difficult.

Overall, IoT-ML systems are transitioning from experimental pilots to operational tools, but widespread adoption will require addressing these systemic barriers.

7. Knowledge Gaps and Challenges

Despite rapid advances in IoT-ML applications for water quality monitoring, significant gaps remain that limit their large-scale adoption and integration into regulatory frameworks. These challenges span technical, methodological, and institutional dimensions, and addressing them is critical for transitioning from experimental pilots to robust operational systems.

7.1. Sensor Reliability and Data Quality

One of the most persistent challenges is the reliability of IoT sensors in field deployments. Biofouling, drift, and calibration issues compromise data accuracy, particularly in nutrient-rich or high-turbidity environments. Many studies report frequent data gaps due to sensor malfunction or communication failures, leading to incomplete time series that undermine ML performance (Huang and Khabusi, 2025). Low-cost sensors are particularly prone to instability, raising concerns about their long-term suitability for compliance monitoring. Development of self-calibrating sensors, advanced anti-fouling materials, and redundancy in sensor networks are needed to improve resilience.

7.2. Data Scarcity and Imbalance

Although IoT systems generate large volumes of data, the availability of labelled datasets for supervised ML

tasks remains limited. For instance, microbial contamination events in bathing waters or intrusion events in drinking water systems are rare, resulting in imbalanced datasets that bias model training (Abdulla and Jameel, 2023). This problem leads to overestimation of accuracy in controlled experiments and poor generalization in operational settings. Data scarcity is further exacerbated in groundwater and coastal systems, where sensors are sparse and costly to deploy (Abdulla and Jameel, 2023). Semi-supervised learning, transfer learning, and anomaly detection methods offer partial solutions but remain underexplored in water quality applications.

7.3. Generalization across Sites and Conditions

Most published studies are site-specific, with models trained and validated in a single catchment, treatment plant, or aquaculture farm. When applied to different sites, performance typically degrades due to variations in hydrology, climate, land use, and infrastructure. Seasonal and interannual variability further complicate model transferability. This lack of generalization limits the scalability of IoT-ML systems for national or regional monitoring. Research into domain adaptation, physics-informed ML, and standardized benchmarking datasets is required to overcome this challenge.

7.4. Integration of Uncertainty and Interpretability

Few studies quantify uncertainty in IoT-ML predictions, despite its importance for regulatory and management decisions. Most outputs are presented as deterministic forecasts without confidence intervals, which limits trust among stakeholders. Similarly, deep learning models are often criticized as "black boxes," hindering interpretability. Without clear explanations of model behavior, regulatory acceptance is unlikely. Incorporating Bayesian methods, ensemble modelling, and explainable AI (XAI) into water quality applications remains a critical research need.

7.5. Cybersecurity and Data Governance

The increasing reliance on IoT networks introduces vulnerabilities to cyberattacks and data manipulation. To date, only a limited number of studies explicitly address cybersecurity in water quality IoT systems. Potential threats include spoofed sensor data, unauthorized access to control systems, and denial-of-service attacks. In addition, the absence of standardized data governance frameworks impedes interoperability between platforms and across jurisdictions. Proprietary systems often restrict data sharing, creating silos that undermine scalability. Development of open standards, encryption protocols, and blockchain-based verification could help safeguard data integrity and foster trust.

7.6. Cost and Scalability Constraints

Although IoT and ML technologies are becoming more affordable, large-scale deployment remains constrained by high capital and maintenance costs. Multi-parameter sondes and spectroscopic sensors remain expensive, and frequent servicing increases operational costs. Low-cost sensor networks, while promising, often sacrifice accuracy and require extensive calibration. Most current deployments are limited to pilot projects, with limited evidence of long-term cost-effectiveness at basin or national scale. Research into hybrid systems that integrate low-cost IoT nodes with high-precision reference sensors could provide a path toward scalable deployment.

7.7. Limited Integration into Regulatory Frameworks

Despite their potential, IoT-ML systems are rarely embedded into formal monitoring programs. Regulatory agencies remain cautious, citing concerns over data reliability, model interpretability, and uncertainty. For example, most bathing water quality advisories are still based on culture-based microbiological analyses despite the availability of IoT-ML forecasting systems. Bridging this gap requires robust validation, inter-agency collaboration, and pilot programs that demonstrate regulatory equivalence or superiority. Until these systems are recognized in environmental directives, their influence will remain limited to research and pilot-scale operations.

7.8. Cross-Domain Synthesis of Challenges

Across domains, several systemic barriers emerge. First, technical reliability of sensors remains the foundation

on which ML performance depends. Second, data challenges—scarcity, imbalance, and lack of standardization—constrain model development and transferability. Third, institutional factors—cybersecurity, governance, cost, and regulatory acceptance—determine whether systems move from pilots to operational use. Without progress in these areas, IoT-ML systems risk remaining experimental technologies rather than mainstream monitoring tools.

8. Future Research Directions and Opportunities

The future of IoT-ML systems in water quality monitoring lies in addressing the persistent technical, methodological, and institutional challenges that currently constrain their widespread adoption. Progress requires coordinated advances across sensor technology, data science, and governance. Beyond incremental improvements, there is an opportunity to fundamentally reimagine how environmental monitoring is designed, implemented, and integrated into decision-making frameworks.

8.1. Advancing Resilient Sensor Technologies

Improving the robustness of sensors is foundational. Many deployments fail not because of ML limitations but due to unreliable sensor data. Research must therefore prioritize anti-fouling innovations, such as bio-inspired coatings, ultraviolet self-cleaning mechanisms, or automated wipers that minimize biofilm accumulation. Equally important is the development of sensors capable of self-calibration, with built-in routines to adjust baselines remotely, reducing the need for frequent site visits. Energy autonomy is another frontier. IoT nodes that integrate solar cells, hydrokinetic turbines, or microbial fuel cells could operate for extended periods without manual recharging, a necessity for remote and inaccessible sites. Hybrid strategies that combine dense grids of low-cost sensors with a few high-quality reference sondes will likely dominate future deployments, striking a balance between affordability and accuracy.

8.2. Expanding Data Availability and Diversity

The lack of representative datasets remains a bottleneck for ML training. Future work must focus on establishing open, standardized repositories of IoT water quality time series. Such benchmark datasets, similar to ImageNet in computer vision, would enable fair comparison of algorithms and accelerate innovation. Generative models and physics-based simulators can also be used to create synthetic datasets that represent rare events, such as contamination intrusions or hypoxic crises, which are underrepresented in real-world data. Participatory sensing offers another avenue. By integrating community-collected data from citizen science initiatives with formal IoT networks, spatial coverage could be greatly expanded. To ensure comparability, standardized metadata protocols such as the OGC Sensor Things API should be adopted widely, making datasets interoperable across regions and platforms (Horsburgh *et al.,* 2025).

8.3. Improving Generalization and Transferability

One of the greatest weaknesses in current studies is the lack of model generalizability across sites and time periods. Most published models are validated within a single catchment or facility, yielding optimistic results that fail to translate elsewhere. Future research should prioritize domain adaptation and transfer learning, allowing models trained in one location to be efficiently retrained in another with minimal data. Embedding physical constraints into ML architectures represents another promising path. Physics-informed neural networks that respect conservation laws and hydrological principles can reduce overfitting and enhance plausibility. Benchmarking protocols also need reform. Rolling-origin evaluations for time series and leave-one-site-out cross-validation should replace random partitioning, providing more realistic assessments of transferability. Meta-learning approaches, in which models learn across multiple catchments or treatment plants, could further enhance adaptability.

8.4. Integrating Uncertainty Quantification and Interpretability

For IoT-ML predictions to be actionable in regulatory settings, they must include quantified uncertainty and transparent reasoning. Current studies overwhelmingly report deterministic point forecasts, a practice that undermines trust. Probabilistic methods such as Bayesian neural networks, ensemble modelling, and quantile regression should be incorporated to provide prediction intervals. Parallel advances are needed in explainable

AI. Tools such as SHAP and LIME can identify which variables most strongly influence predictions, turning black-box models into interpretable systems. For instance, recent work on bathing water forecasts revealed that rainfall and turbidity were consistently the dominant drivers when explainability frameworks were applied. Future decision-support systems should embed both uncertainty and interpretability, allowing regulators to weigh predictions in light of risk tolerance and management thresholds.

8.5. Strengthening Cybersecurity and Governance

As IoT monitoring scales, the risk of cyber-physical vulnerabilities becomes increasingly significant. Few studies explicitly address how to secure water IoT networks against spoofed data, denial-of-service attacks, or unauthorized control. End-to-end encryption tailored to low-power IoT devices should become standard practice, while blockchain and distributed ledger systems offer promise for ensuring tamper-proof audit trails. Governance must advance alongside technical security. Proprietary platforms currently create silos that limit interoperability and data sharing. International frameworks for water data governance, including standardized metadata, security compliance protocols, and interoperability guidelines, will be essential for cross-border and national-scale systems.

8.6. Enhancing Cost-Effectiveness and Sustainability

Economic and environmental sustainability must be integral to future IoT-ML designs. Large-scale deployments will only be feasible if life-cycle costs, including maintenance, calibration, and replacement, are minimized. Rigorous cost–benefit analyses comparing IoT–ML with conventional monitoring are needed, not only in terms of financial savings but also in avoided pollution incidents, reduced energy use, and improved regulatory responsiveness. Environmental sustainability is equally critical. Few studies consider the ecological footprint of sensor networks, including battery waste and rare-earth material demand. Future IoT designs should prioritize modular, repairable, and recyclable devices. Linking water quality innovation with circular economy principles would ensure that scaling IoT does not introduce new environmental burdens.

8.7. Pathways to Regulatory Integration

Perhaps the most decisive frontier is regulatory acceptance. Despite promising results, IoT-ML systems are rarely embedded in official compliance frameworks. To build credibility, large-scale pilot programs must be undertaken where IoT-ML predictions directly inform bathing water advisories, wastewater discharge permits, or drinking water safety protocols. Inter-laboratory validation studies comparing IoT-ML forecasts with traditional laboratory results are essential to demonstrate equivalence. Legal and liability frameworks must also evolve, clarifying responsibility when automated predictions guide interventions. Finally, capacity building is vital. Regulators and utility operators will need training not only in sensor maintenance but also in interpreting ML outputs, uncertainty intervals, and data dashboards.

8.8. Cross-Cutting Opportunities

Beyond addressing specific challenges, several transformative opportunities lie ahead. One is the integration of IoT-ML with Earth observation data from satellites and radar, enabling multi-scale monitoring that links basin-level drivers with in situ dynamics. Another is the rise of digital twins for water systems, dynamic virtual replicas of rivers, lakes, or treatment plants that update continuously with IoT and ML inputs. These tools could allow scenario testing and real-time management interventions. Finally, IoT-ML can converge with emerging technologies such as autonomous drones for water sampling, blockchain for transparent reporting, and climate models for adaptive management. These cross-cutting opportunities position IoT-ML not merely as incremental monitoring tools but as central pillars in the digital transformation of water resource management.

8.9. Synthesis

The trajectory of IoT-ML research indicates a clear transition from isolated pilots to integrated, large-scale deployments. Achieving this transformation requires simultaneous advances in resilient sensor hardware, methodological rigor, and governance frameworks. Progress in one dimension without the others will be insufficient. A future in which IoT-ML systems are trusted components of regulatory infrastructure is possible,

but only if uncertainty, interpretability, sustainability, and security are prioritized alongside accuracy. If realized, such systems could fundamentally change how societies monitor, predict, and protect water resources in an era of increasing environmental stress.

9. Conclusion and Policy Implications

The integration of machine learning and Internet of Things technologies into water quality monitoring represents a significant shift in how aquatic environments are observed and managed. Evidence from diverse domains—rivers, lakes, groundwater, drinking water networks, wastewater treatment plants, bathing waters, aquaculture, and coastal systems—demonstrates the ability of IoT-ML systems to provide continuous, high-resolution data and predictive insights that far exceed the capabilities of traditional monitoring. These advances can improve responsiveness to contamination events, optimize operational decisions, and support adaptive management under conditions of increasing environmental stress. At the same time, this review highlights critical barriers that prevent widespread adoption. Technical challenges include sensor biofouling, calibration drift, data gaps, and energy supply limitations. Methodological issues persist around data scarcity, generalization across sites, lack of uncertainty quantification, and limited interpretability of deep learning models. Institutional barriers—particularly high costs, fragmented governance, cybersecurity risks, and the absence of regulatory frameworks—remain decisive in determining whether IoT-ML systems can transition from pilot projects to mainstream monitoring tools.

The policy implications are clear. Regulatory agencies should begin to recognize the value of IoT-ML forecasts as complements, and eventually partial replacements, for conventional laboratory analyses. This will require structured pilot programs, inter-laboratory validation, and the development of legal frameworks for accountability in automated decision-making. Public-private partnerships may be essential to finance large-scale infrastructures, while international cooperation will be needed to standardize data formats, metadata, and interoperability protocols. Capacity building among regulators and utility staff must also be prioritized to ensure that the technical and interpretive skills required for IoT-ML adoption are widely available.

Looking forward, the convergence of resilient sensor hardware, advanced ML architectures, and robust governance frameworks offers an opportunity to transform water quality monitoring from a reactive to a proactive discipline. If the research priorities identified in this review are addressed, IoT-ML systems could become trusted regulatory tools, enabling real-time protection of public health, safeguarding of ecosystems, and more efficient management of water resources. Achieving this vision will demand sustained collaboration between scientists, engineers, policymakers, and industry. The stakes are high: with climate change, population growth, and pollution pressures intensifying, the capacity to predict and prevent water quality crises is no longer optional but essential.

Disclaimer (Artificial Intelligence)

Author hereby declares that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Acknowledgment

Competing Interests

Author has declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

Authors' Contributions

The sole author designed, analyzed, interpreted and prepared the manuscript.

References

Abdulla, A.R. and M. Jameel, N.G. (2023). A Review on IoT Intrusion Detection Systems Using Supervised Machine Learning: Techniques, Datasets, and Algorithms. *UHD Journal of Science and Technology*, 7(1), 53-65. https://doi.org/10.21928/uhdjst.v7n1y2023.pp53-65

Abdullah, A.F., Man, H.C., Mohammed, A., Karim, M.M.A., Yunusa, S.U. and Jais, N.A.B.M. (2024). Charting the Aquaculture Internet of Things Impact: Key Applications, Challenges, and Future Trend. *Aquaculture Reports*, 39, 102358. https://doi.org/10.1016/j.aqrep.2024.102358

Abo, L.D., Areti, H.A., Jayakumar, M., Rangaraju, M. and Subashini, S. (2025). Nanobiomaterials-Enabled Sensors for Heavy Metal Detection and Remediation in Wastewater Systems: Advances in Synthesis, Characterization, and Environmental Applications. *Results in Engineering*, 27, 105694. https://doi.org/10.1016/j.rineng.2025.105694

Adelagun, R.O.A., Edet Etim, E. and Emmanuel Godwin, O. (2021). Application of Water Quality Index for the Assessment of Water from Different Sources in Nigeria. *Promising Techniques for Wastewater Treatment and Water Quality Assessment*. https://doi.org/10.5772/intechopen.98696

Ahmed, A.A., Sayed, S., Abdoulhalik, A., Moutari, S. and Oyedele, L. (2024). Applications of Machine Learning to Water Resources Management: A Review of Present Status and Future Opportunities. *Journal of Cleaner Production*, 441, 140715. https://doi.org/10.1016/j.jclepro.2024.140715

Ahmed, S.F., Shanjana Shuravi Shawon, Shaila Afrin, Rafa, S.J., Hoque, M. and Gandomi, A.H. (2025). Optimising Internet of Things (IoT) Performance Through Cloud, Fog and Edge Computing Architecture. *IET Wireless Sensor Systems*, 15(1). https://doi.org/10.1049/wss2.70016

Ali, O., Ishak, M.K. and Bhatti, M.K.L. (2021). Emerging IoT Domains, Current Standings and Open Research Challenges: A Review. *PeerJ Computer Science*, 7, e659. https://doi.org/10.7717/peerj-cs.659

Baena-Navarro, R., Carriazo-Regino, Y., Torres-Hoyos, F. and Pinedo-López, J. (2025). Intelligent Prediction and Continuous Monitoring of Water Quality in Aquaculture: Integration of Machine Learning and Internet of Things for Sustainable Management. *Water*, 17(1), 82. https://doi.org/10.3390/w17010082

Brenckman, C.M., Parameswarappa Jayalakshmamma, M., Pennock, W.H., Ashraf, F. and Borgaonkar, A.D. (2025). A Review of Harmful Algal Blooms: Causes, Effects, Monitoring, and Prevention Methods. *Water*, 17(13), 1980. https://doi.org/10.3390/w17131980

Busari, I., Sahoo, D., Harmel, R.D. and Haggard, B.E. (2023). A Review of Machine Learning Models for Harmful Algal Bloom Monitoring in Freshwater Systems. *Journal of Natural Resources and Agricultural Ecosystems*, 1(2), 63-76. https://doi.org/10.13031/jnrae.15647

Caballero, C.B., Martins, V.S., Paulino, R.S., Butler, E., Sparks, E., Lima, T.M. and Novo, E.M.L.M. (2025). The Need for Advancing Algal Bloom Forecasting Using Remote Sensing and Modelling: Progress and Future Directions. *Ecological Indicators*, 172, 113244. https://doi.org/10.1016/j.ecolind.2025.113244

Cao, C., Debnath, R. and Alvarez, R.M. (2025). Physics-Based Machine Learning for Predicting Urban Air Pollution Using Decadal Time Series Data. *Environmental Research Communications*. https://doi.org/10.1088/2515-7620/add795

Chaulagain, S., Lamichhane, M. and Chaulagain, U. (2025). A Review of Current Trends, Challenges, and Future Perspectives in Machine Learning Applications to Water Resources in Nepal. *Journal of Hazardous Materials Advances*, 18, 100678. https://doi.org/10.1016/j.hazadv.2025.100678

Crawford, C. (2024). Protocol Power: Matter, IoT Interoperability, and a Critique of Industry Self-Regulation. *Internet Policy Review*, 13(2). https://doi.org/10.14763/2024.2.1776

Delgado, A., Briciu-Burghina, C. and Regan, F. (2021). Antifouling Strategies for Sensors Used in Water Monitoring: Review and Future Perspectives. *Sensors*, 21(2), 389. https://doi.org/10.3390/s21020389

Dharmarathne, G., Abekoon, A.M.S.R., Bogahawaththa, M., Alawatugoda, J. and Meddage, D.P.P. (2025). A Review of Machine Learning and Internet-of-Things on the Water Quality Assessment: Methods, Applications and Future Trends. *Results in Engineering*, 26, 105182. https://doi.org/10.1016/j.rineng.2025.105182

Diane, A., Diallo, O. and El. (2025). A Systematic and Comprehensive Review on Low Power Wide Area Network: Characteristics, Architecture, Applications and Research Challenges. *Discover Internet of Things*, 5(1). https://doi.org/10.1007/s43926-025-00097-6

Ding, S., Ward, H. and Tukker, A. (2023). How Internet of Things Can Influence the Sustainability Performance of Logistics Industries—A Chinese Case Study. *Cleaner Logistics and Supply Chain*, 6, 100094. https://doi.org/10.1016/j.clscn.2023.100094

Dritsas, E. and Trigka, M. (2025). Remote Sensing and Geospatial Analysis in the Big Data Era: A Survey. *Remote Sensing*, 17(3), 550-550. https://doi.org/10.3390/rs17030550

Erun. (2025). Multiparameter Sondes: Essential Tools for Water Quality Monitoring. *Erunwas.com*. https://www.erunwas.com/news-detail/id-164.html

Essamlali, I., Nhaila, H. and Khaili, M.E. (2024). Advances in Machine Learning and IoT for Water Quality Monitoring: A Comprehensive Review. *Heliyon*, 10(6), e27920–e27920. https://doi.org/10.1016/j.heliyon.2024.e27920

Ferdowsi, A., Piadeh, F., Behzadian, K., Mousavi, S.-F. and Ehteram, M. (2024). Urban Water Infrastructure: A Critical Review on Climate Change Impacts and Adaptation Strategies. *Urban Climate*, 58, 102132. https://doi.org/10.1016/j.uclim.2024.102132

Furrer, V., Mutzner, L., Singer, H. and Ort, C. (2023). Micropollutant Concentration Fluctuations in Combined Sewer Overflows Require Short Sampling Intervals. *Water Research X*, 21, 100202-100202. https://doi.org/10.1016/j.wroa.2023.100202

García, J., Leiva-Araos, A., Diaz-Saavedra, E., Moraga, P., Pinto, H. and Yepes, V. (2023). Relevance of Machine Learning Techniques in Water Infrastructure Integrity and Quality: A Review Powered by Natural Language Processing. *Applied Sciences*, 13(22), 12497. https://doi.org/10.3390/app132212497

Hasan, F., Nassereldin Ahmed Kabashi, Saleh, T., Alam, M.Z., Wahab, M.F. and Nour, A.H. (2024). Water Quality Monitoring Using Machine Learning and Iot: A Review. 8(2), 32-54. https://doi.org/10.31436/cnrej.v8i2.100

Hassani, H., Silva, E.S., Combe, M., Andreou, D., Ghodsi, M., Yeganegi, M.R. and Gozlan, R.E. (2019). A Support Vector Machine Based Approach for Predicting the Risk of Freshwater Disease Emergence in England. *Stats*, 2(1), 89-103. https://doi.org/10.3390/stats2010007

Himeur, Y., Ghanem, K., Alsalemi, A., Bensaali, F. and Amira, A. (2021). Artificial Intelligence Based Anomaly Detection of Energy Consumption in Buildings: A Review, Current Trends and New Perspectives. *Applied Energy*, 287(1), 116601. https://doi.org/10.1016/j.apenergy.2021.116601

Horsburgh, J.S., Lippold, K. and Slaugh, D.L. (2025). Adapting OGC's SensorThings API and Data Model to Support Data Management and Sharing for Environmental Sensors. *Environmental Modelling & Software*, 183, 106241. https://doi.org/10.1016/j.envsoft.2024.106241

Hou, Y., Liu, Z., Huang, H., Lou, C., Sun, Z., Liu, X., Pang, J., Ge, S., Wang, Z., Zhou, W. and Liu, H. (2024). Biosensor Based Microfluidic Platforms for Rapid Clinical Detection of Pathogenic Bacteria. *Advanced Functional Materials*. https://doi.org/10.1002/adfm.202411484

Huang, Y.-P. and Khabusi, S.P. (2025). Artificial Intelligence of Things (AIoT) Advances in Aquaculture: A Review. *Processes*, 13(1), 73-73. https://doi.org/10.3390/pr13010073

Igor Gulshin and Kuzina, O. (2024). Optimization of Wastewater Treatment Through Machine Learning-Enhanced Supervisory Control and Data Acquisition: A Case Study of Granular Sludge Process Stability and Predictive Control. *Automation*, 6(1), 2-2. https://doi.org/10.3390/automation6010002

Jayaraman, P., Kothalam Krishnan Nagarajan, Pachaivannan Partheeban and Krishnamurthy, V. (2024). Critical Review on Water Quality Analysis Using IoT and Machine Learning Models. *International Journal of Information Management Data Insights*, 4(1), 100210-100210. https://doi.org/10.1016/j.jjimei.2023.100210

Jørgensen, B.N., Gunasekaran, S.S. and Ma, Z.G. (2025). Impact of EU Laws on AI Adoption in Smart Grids: A Review of Regulatory Barriers, Technological Challenges, and Stakeholder Benefits. *Energies*, 18(12), 3002. https://doi.org/10.3390/en18123002

Jun, M.-J. (2021). A Comparison of a Gradient Boosting Decision Tree, Random Forests, and Artificial Neural Networks to Model Urban Land Use Changes: The Case of the Seoul Metropolitan Area. *International Journal of Geographical Information Science*, 35(11), 2149-2167. https://doi.org/10.1080/13658816.2021.1887490

Karlson, B., Andersen, P., Arneborg, L., Cembella, A., Eikrem, W., John, U., West, J.J., Klemm, K., Kobos, J., Lehtinen, S., Lundholm, N., Mazur-Marzec, H., Naustvoll, L., Poelman, M., Provoost, P., De Rijcke, M. and Suikkanen, S. (2021). Harmful Algal Blooms and their Effects in Coastal Seas of Northern Europe. *Harmful Algae*, 102, 101989. https://doi.org/10.1016/j.hal.2021.101989

Katie, B. (2024). Internet of Things (IoT) for Environmental Monitoring. *International Journal of Computing and Engineering*, 6(3), 29-42. https://doi.org/10.47941/ijce.2139

Lombardo, A., Parrino, S., Peruzzi, G. and Pozzebon, A. (2021). LoRaWAN vs NB-IoT: Transmission Performance Analysis within Critical Environments. *IEEE Internet of Things Journal*, 1-1. https://doi.org/10.1109/jiot.2021.3079567

Madrid, Y. and Zayas, Z.P. (2007). Water Sampling: Traditional Methods and New Approaches in Water Sampling Strategy. *TrAC Trends in Analytical Chemistry*, 26(4), 293-299. https://doi.org/10.1016/j.trac.2007.01.002

Ngwenya, B., Paepae, T. and Bokoro, P.N. (2025). Monitoring Ambient Water Quality Using Machine Learning and IoT: A Review and Recommendations for Advancing SDG Indicator 6.3.2. *Journal of Water Process Engineering*, 73, 107664. https://doi.org/10.1016/j.jwpe.2025.107664

Okafor, N. (2023). Advances and Challenges in IoT Sensors Data Handling and Processing in Environmental Monitoring Systems. https://doi.org/10.36227/techrxiv.24045951.v1

Okafor, N., Ingle, R., Matthew, U.O., Saunders, M. and Delaney, D.T. (2024). Assessing and Improving IoT Sensor Data Quality in Environmental Monitoring Networks: A Focus on Peatlands. *IEEE Internet of Things Journal*, 11(24), 40727-40742. https://doi.org/10.1109/jiot.2024.3454241

Pang, Z., Zhou, Z., Fu, J., Jiang, W., Qin, X. and Sun, M. (2025). Deep Learning-Based Remote Sensing Retrieval of Inland Water Quality: A Review. *Journal of Hydrology: Regional Studies*, 61, 102759. https://doi.org/10.1016/j.ejrh.2025.102759

Pires, L.M. and Gomes, J. (2024). River Water Quality Monitoring Using LoRa-Based IoT. *Designs*, 8(6), 127. https://doi.org/10.3390/designs8060127

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S. and Pélissier, R. (2020). Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models. *Nature Communications*, 11(1), 4540. https://doi.org/10.1038/s41467-020-18321-y

Politikos, D.V., Petasis, G. and Katselis, G. (2021). Interpretable Machine Learning to Forecast Hypoxia in a Lagoon. *Ecological Informatics*, 66, 101480. https://doi.org/10.1016/j.ecoinf.2021.101480

R, S.A. and Jeevaa Katiravan. (2025). Enhancing Anomaly Detection and Prevention in Internet of Things (IoT) Using Deep Neural Networks and Blockchain Based Cyber Security. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-04164-4

Rahman, Md. S., Ghosh, T., Aurna, N.F., Kaiser, M.S., Anannya, M. and Hosen, A.S.M.S. (2023). Machine Learning and Internet of Things in Industry 4.0: A Review. *Measurement: Sensors*, 28, 100822. https://doi.org/10.1016/j.measen.2023.100822

Rastegari, H., Nadi, F., Lam, S.S., Ikhwanuddin, M., Kasan, N.A., Rahmat, R.F. and Mahari, W.A.W. (2023). Internet of Things in Aquaculture: A Review of the Challenges and Potential Solutions Based on Current and Future Trends. *Smart Agricultural Technology*, 4, 100187. https://doi.org/10.1016/j.atech.2023.100187

Roy, S.C., Islam, M.A., Sarkar, R., Sarkar, R.R., Jibon, F.A. and Naznin, L. (2025). A Study of Water Quality Monitoring System with Internet of Things and Machine Learning Regression Techniques. *Cureus Journal of Computer Science*. https://doi.org/10.7759/s44389-025-04028-y

Saeik, F., Avgeris, M., Spatharakis, D., Santi, N., Dechouniotis, D., Violos, J., Leivadeas, A., Athanasopoulos, N., Mitton, N. and Papavassiliou, S. (2021). Task Offloading in Edge and Cloud Computing: A Survey on Mathematical, Artificial Intelligence and Control Theory Solutions. *Computer Networks*, 195, 108177. https://doi.org/10.1016/j.comnet.2021.108177

Shah, N.W., Baillie, B.R., Bishop, K., Ferraz, S., Högbom, L. and Nettles, J. (2022). The Effects of Forest Management on Water Quality. *Forest Ecology and Management*, 522(120397), 120397. https://doi.org/10.1016/j.foreco.2022.120397

Shaharuddin, S., Abdul Maulud, K.N., Syed Abdul Rahman, S.A.F., Che Ani, A.I. and Pradhan, B. (2023). The Role of IoT Sensor in Smart Building Context for Indoor Fire Hazard Scenario: A Systematic Review of Interdisciplinary Articles. *Internet of Things*, 22, 100803. https://doi.org/10.1016/j.iot.2023.100803

Shukla, B.K., Ruchi Saraswat, Agarwal, N., Singh, H.K. and Verma, S. (2025). A Comparative Study of IoT-Based Water Quality Monitoring Systems (IoT-WQMS) and the Potential of Machine Learning in Water Quality Assessment. 1-23. https://doi.org/10.1007/978-981-96-3886-4_1

Singh, N., Buyya, R. and Kim, H. (2024). Securing Cloud-Based Internet of Things: Challenges and Mitigations. *Sensors*, 25(1), 79-79. https://doi.org/10.3390/s25010079

Staradumskytë, D. and Paulauskas, A. (2012). Indicators of Microbial Drinking and Recreational Water Quality. *Biologija*, 58(1). https://doi.org/10.6001/biologija.v58i1.2317

Truong, A.M. and Luong, H.Q. (2024). A Non-Destructive, Autoencoder-Based Approach to Detecting Defects and Contamination in Reusable Food Packaging. *Current Research in Food Science*, 8, 100758. https://doi.org/10.1016/j.crfs.2024.100758

Wang, A., Li, H., He, Z., Tao, Y., Wang, H., Yang, M., Savic, D., Daigger, G.T. and Ren, N. (2024). Digital Twins for Wastewater Treatment: A Technical Review. *Engineering*. https://doi.org/10.1016/j.eng.2024.04.012

Yang, S., Behzadian, K., Coleman, C., Holloway, T.G. and Campos, L.C. (2025). Application of AI-Based Techniques for Anomaly Management in Wastewater Treatment Plants: A Review. *Journal of Environmental Management*, 392, 126886. https://doi.org/10.1016/j.jenvman.2025.126886

Zhao, X., Wang, H., Bai, M., Xu, Y., Dong, S., Rao, H. and Ming, W. (2024). A Comprehensive Review of Methods for Hydrological Forecasting Based on Deep Learning. *Water*, 16(10), 1407-1407. https://doi.org/10.3390/w16101407

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B. and Ye, L. (2022). A Review of the Application of Machine Learning in Water Quality Evaluation. *Eco-Environment & Health*, 1(2). https://doi.org/10.1016/j.eehl.2022.06.001

Zulkifli, C.Z., Garfan, S., Talal, M., Alamoodi, A.H., Alamleh, A., Ahmaro, I.Y.Y., Sulaiman, S., Ibrahim, A.B., Zaidan, B.B., Ismail, A.R., Albahri, O.S., Albahri, A.S., Soon, C.F., Harun, N.H. and Chiang, H.H. (2022). IoT-Based Water Monitoring Systems: A Systematic Review. *Water*, 14(22), 3621. https://doi.org/10.3390/w14223621