Machine Learning Algorithm for Water Quality Classification: A Systematic Literature Review

Daniel Idzwan Darwis[1], M. Shahkhir Mozamir[1]*, Najwan Khambari[1], Norharyati Harum[2], Izzatie Husna Fauzi[3], Anggi Muhammad Rifa'i[4]

[1]Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Malaysia

[2]Smart Manufacturing Technology Centre (SMTC), Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal Melaka, Malaysia

[3]Faculty of Computing, University Malaysia Pahang Al-Sultan Abdullah, Pekan 26600, Pahang, Malaysia

[4]Department of Informatic Engineering, Faculty of Engineering, Universitas Pelita Bangsa, Bekasi-Jawa Barat, Indonesia

***Correspondence Author**

DOI: https://doi.org/10.47772/IJRISS.2026.10100291

Received: 16 January 2026; Accepted: 21 January 2026; Published: 03 February 2026

ABSTRACT

Assessing Water quality classification has become an important research area as the demand for clean and safe water continues to grow worldwide. In recent years, Machine Learning (ML) has shown great potential in improving how water quality is monitored and analyzed. By using ML models, researchers can process large and complex environmental data more effectively to detect pollution, predict water conditions, and support better management decisions. While many studies have focused on using sensors and data analytics for monitoring, only a few have provided a full review of the different ML methods and their effectiveness in classifying water quality. Therefore, this paper aims to achieve two main goals: (1) to conduct a Systematic Literature Review (SLR) of existing ML techniques applied in water quality classification, and (2) to identify the main findings, challenges, and future opportunities in this field. Through a careful review and comparison of previous research, this paper hopes to give a clearer overview of how ML contributes to water quality analysis and guide future work in creating more accurate and intelligent systems for real-world environmental applications.

Keywords: Machine Learning, Water Quality Monitoring, Classification, Environmental Management, Smart Sensing Systems

INTRODUCTION

Water is a fundamental resource for life, yet its quality is increasingly threatened by pollution, climate change, and infrastructure challenges. Traditional water-quality assessment methods such as laboratory sampling and manual chemical testing are reliable, but they are time-consuming, expensive, and unsuitable for continuous real-time monitoring. Recent reviews emphasize that the volume of available aquatic-environment data has grown substantially, making advanced data-driven methods essential for efficient classification and prediction of water quality [1].

In parallel, machine learning (ML) techniques have emerged as promising tools for water quality classification, capable of handling large datasets, recognizing complex nonlinear relationships, and improving classification accuracy beyond what conventional models can offer. Studies have applied a range of ML models from Support Vector Machines and Random Forests to Deep Neural Networks and ensemble frameworks to classify water quality based on multiple SENSOR inputs and environmental variables [2]. Despite these advances, the applicability of such models across different water environments remains fragmented, with variations in parameter selection, model interpretability and data-quality issues limiting broader deployment.

Recognizing this gap, this paper presents a Systematic Literature Review (SLR) of research published between 2019 and 2025 focused on machine-learning methods for water quality classification. The objectives of this review are two-fold: (1) to identify, categorize and compare machine learning models and their performance in water quality classification tasks; and (2) to highlight key challenges, limitations and future research directions in the field. By synthesizing findings from multiple studies, this review aims to provide insights for researchers seeking to develop robust, generalizable and real-time water quality monitoring systems.

Motivation & Related Work

Several factors motivated the authors to conduct this systematic literature review (SLR). While previous studies on machine learning for water quality classification exist, most have not provided an in-depth discussion of the specific processes and techniques involved in data analysis and model development. Based on the study conducted by Yang et al. the research highlights the increasing concern over water pollution, which poses serious risks to aquatic life, human health, and overall environmental sustainability. As the demand for clean and safe water continues to grow, traditional monitoring methods are often found to be inefficient, time-consuming, and unable to provide real-time analysis. The authors identified limitations in current approaches, particularly in terms of data resolution, atmospheric correction, and the general applicability of models for accurate water quality assessment. To address these issues, the study explores the potential of machine learning (ML) as a modern solution for classifying and predicting water quality parameters more efficiently. By leveraging algorithms such as Support Vector Machine (SVM), Random Forest, and Artificial Neural Networks (ANNs), machine learning offers improved accuracy and adaptability across various environmental conditions [3].

Additionally, the integration of remote sensing and IoT-based sensors has further enhanced data collection and real-time monitoring capabilities. Overall, this research contributes valuable insights to the field of environmental monitoring by demonstrating how machine learning can revolutionize water quality classification, supporting sustainable water management practices and enabling faster, more reliable decision-making processes for the protection of water resources.

Moreover, based on the study conducted by the authors Lokman et al., the research emphasizes the urgent need for effective prediction and management of water quality, particularly in countries like Malaysia, where rapid industrialization, agricultural runoff, and urban development have contributed significantly to water pollution. This issue is critical because maintaining clean and sustainable water resources is essential for both environmental balance and public health. The authors identified limitations in existing predictive models, particularly in terms of data quality, model interpretability, and the integration of spatio-temporal and fuzzy logic techniques, which often hinder accurate water quality assessment. To overcome these challenges, the study focuses on the potential of machine learning (ML) approaches to enhance forecasting and classification performance. By analyzing various ML models such as Support Vector Machines (SVM), Random Forests (RF), Artificial Neural Networks (ANNs), and hybrid models the research aims to identify the most efficient and interpretable algorithms for improving prediction accuracy. Overall, this paper contributes to the growing body of work in environmental informatics by providing a comprehensive evaluation of how machine learning techniques can advance water quality monitoring and support sustainable water management strategies in polluted and high-risk regions [1].

Next, based on the study conducted by the Talukdar et al., the research underscores the growing importance of water quality monitoring and modeling as a means to protect and improve aquatic ecosystems and their surrounding environments. Ensuring clean and safe water resources is vital not only for ecological sustainability but also for public health and socio-economic development. The study highlights the integration of computer science and mathematical modeling in developing advanced algorithms capable of accurately assessing and predicting water quality parameters an essential step toward effective environmental management. However, the authors identified a significant gap in existing research, particularly in the comparative evaluation of different water quality models and their effectiveness across diverse environmental conditions. To address this issue, the paper aims to review and analyze various machine learning and statistical models used for water quality classification, providing insights into their strengths, limitations, and suitability for different scenarios. Overall, this research contributes to the broader understanding of data-driven water quality management by offering practical guidance for selecting and applying appropriate computational tools in environmental monitoring and decision-making [4].

Based on the study conducted by Ejigu, the research focuses on the increasing challenges of water quality modeling caused by rising water demands driven by population growth, urbanization, and industrialization, which collectively contribute to the deterioration of water resources. Ensuring clean and sustainable water systems has become a critical global priority, as poor water quality directly impacts human health, ecosystems, and economic activities. The study emphasizes the importance of water quality modeling as a fundamental tool for effective management, monitoring, and policy formulation in integrated water resource and environmental management frameworks. However, the authors identified a notable gap in the standardization and harmonization of modeling approaches, as existing models often differ in structure, parameterization, and applicability. To address this issue, the paper aims to review and compare major water quality models, assessing their strengths, limitations, and inherent uncertainties. Overall, the study contributes valuable insights toward improving the accuracy, reliability, and adaptability of water quality models, supporting better decision-making and sustainable water management practices [5].

Lastly, based on the study conducted by Yan et al, the research focuses on the urgent need for effective surface water quality monitoring and management, recognizing its importance for essential human activities such as agriculture, industry, and daily consumption. Maintaining high-quality water resources is crucial for ensuring public health, ecological balance, and sustainable development. The study emphasizes the significance of water quality indices (water quality classificationIs, TSIs, HMIs) as key indicators for assessing and managing water quality conditions. However, the authors identified a major gap in the integration of advanced technologies, particularly those that enable real-time monitoring and predictive management systems.

To address this issue, the paper proposes the development of a next-generation water quality management framework that leverages expert systems and machine learning algorithms to enhance the precision, speed, and reliability of water quality assessments. Overall, this research contributes to the advancement of intelligent environmental management systems, paving the way for smarter and more sustainable water resource monitoring in the future [6].

METHODOLOGY

To structure the review process effectively, this study adopts the Systematic Literature Review (SLR) methodology. The research framework illustrated in Figure 1 is developed based on the SLR guidelines proposed by [7]. The methodology is divided into three main stages: (1) Preparation, (2) Organization, and (3) Results and Discussion. Each stage consists of several systematic steps. In the Preparation stage, the tasks include recognizing the need for the review and formulating the research questions. The Organization stage covers key activities such as the search process, screening of relevant studies, and data extraction and synthesis. Finally, the Results and Discussion stage focuses on analyzing, interpreting, and presenting the findings in detail. The specific implementation and sub-stages of each phase are further elaborated in the following sections.

The discussion section shows how the author interprets the results considering what was already known, and to explain the new understanding of the problem after taking your results into consideration. The discussion must connect with the Introduction, so it tells how your study contributes to the body of knowledge and society.

INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN SOCIAL SCIENCE (IJRISS)
ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue I January 2026

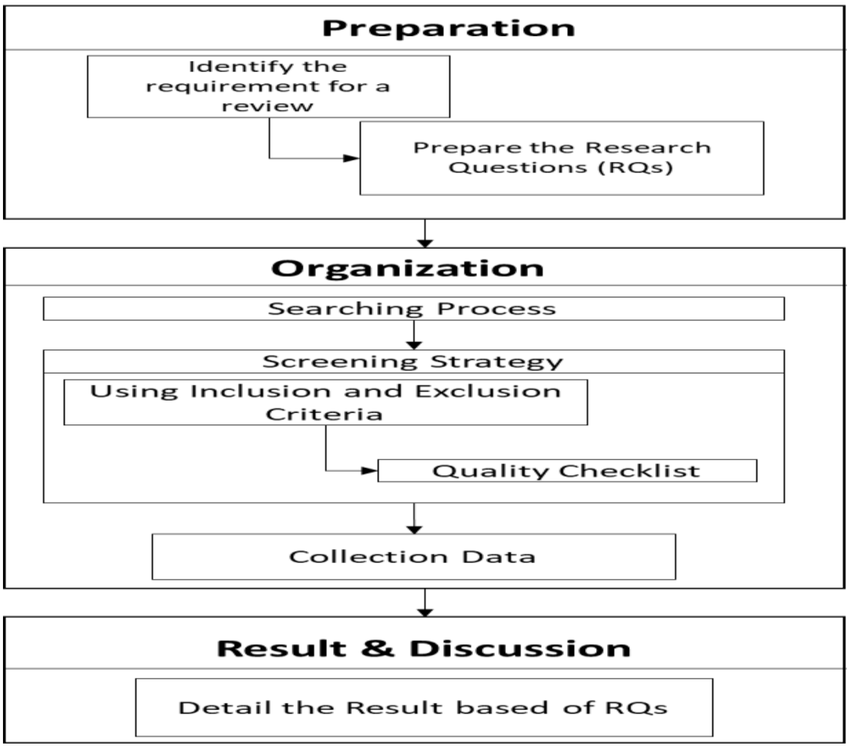

Figure 1. Review Research Methodology

Preparation Phase

The justification for conducting this review is explained in the previous section (Motivation and Related Works), which examined earlier studies and review papers related to water quality assessment. From the analysis, it was found that many existing SLRs lack detailed discussions on the machine learning algorithms currently used for water quality classification and prediction. Therefore, this SLR was developed to address that gap by providing a comprehensive analytical review of different algorithms highlighting their strengths, limitations, processes, and overall effectiveness in handling various challenges related to water quality monitoring. The methodological framework and research questions of this SLR were formulated based on the study's main objective to enhance understanding and application of machine learning in this domain. The following Research Questions (RQs) have been formulated, and the rationale for each one of them is provided in Table 1.

Table 1. Research Questions (RQs)

Research Questions (RQs)	Motivations
RQ1: What types of papers are covered by the investigation?	To identify the different sets of the finding studies in the domain.
RQ2: What are the most commonly used or compared machine learning algorithms for classifying and predicting water quality?	To identify the algorithms that have been most frequently applied or compared for water quality classification.
RQ3: How many sensors are used, and what types are implemented for water quality classification?	To determine the number and types of sensors typically used in water quality classification systems.
RQ4: What are the performance metrics used to evaluate classification algorithms in water quality classification?	To evaluate the performance of classification algorithms in water quality classification is essential to determine their effectiveness, reliability, and applicability.

RQ5: What kind of research on water quality classification using machine learning has been conducted?	To determine the field of research related to water quality classification using machine learning.
RQ6: What are the key challenges and limitations associated with classification algorithms in water quality classification using machine learning?	To determine the challenges and limitations in applying machine learning algorithms for water quality classification.
RQ7: What are the future trends and potential research directions for machine learning algorithms in water quality classification?	To identify new trends, future research directions, and innovative ideas in the field of machine learning for water quality classification.

Organization Phase

This phase involves the execution of the specified stages: a systematic methodology for pinpointing pertinent studies, the approach utilized for selecting the articles to be included, and the process of data acquisition and analysis. The subsequent subsections will elaborate on the implementation of each of these stages below:

Searching Process

The proper definition of the search process is essential to ensure accurate and reliable outcomes [8] In this SLR, the identification and selection of sources are conducted systematically to gather all relevant studies related to machine learning for water quality classification. This process is guided by two (2) main elements: 1) the use of diverse and well-structured search strings, and 2) the selection of credible and appropriate academic databases.

The search terms are formulated in this SLR based on the listed Research Questions (RQs) and standard procedure in which consists of the following steps. In the formulation of the search terms of this SLR, it has its basis on the listed research questions as well as the standard procedure that entails the following steps [1]. Below are the steps:

1) Recognizing the related keywords of these Research Questions (RQs).

2) Including the expected synonyms and spelling of the term's alternatives.

3) Checking the appropriateness of the study search terms.

4) Compiling the search terms with Boolean OR or AND operators.

The result for the search strings after following the steps above as below:

- ("Machine Learning" OR "Artificial Intelligence") AND "Water Quality Classification")
- ("Water Quality Prediction" OR "Water Quality Assessment" OR "Water Quality Monitoring") AND ("Machine Learning" OR "Deep Learning")
- ("Water Quality") AND ("Classification" OR "Prediction" OR "Detection") AND ("Neural Network" OR "Support Vector Machine" OR "Random Forest")
- ("Machine Learning Models" OR "AI Models") AND ("Water Quality Parameters" OR "Water Pollution")
- ("Machine Learning" OR "Artificial Intelligence") AND ("Water Quality Classification") AND ("Supervised Learning" OR "Unsupervised Learning" OR "Hybrid Models")

In this SLR, the search process was conducted using three major electronic databases: Scopus, IEEE Xplore, and ScienceDirect. These databases were selected because of their extensive coverage of high-quality research papers, technical reports, and review articles related to Machine Learning and Water Quality Classification. First, Scopus serves as one of the largest multidisciplinary databases, indexing peer-reviewed journals, books, and conference proceedings across scientific and engineering disciplines [9]. Second, IEEE Xplore, managed by the Institute of Electrical and Electronics Engineers (IEEE), provides access to a wide collection of publications focusing on computer science, engineering innovations, and artificial intelligence applications [10]. Lastly, ScienceDirect, operated by Elsevier, is a comprehensive full-text database offering scientific and technical research materials in environmental science, data analytics, and machine learning [11]. To ensure the relevance and reliability of the findings, this review limits the search to studies published within the last five years, from 2020 to December 2025, capturing the most recent trends and developments in the field.

Screening Strategy

Another key component defined in the research methodology is the study selection strategy, which ensures that only the most relevant and high-quality studies are included in this review. After implementing the defined search process, a total of 60 papers were initially retrieved. To refine these results and identify the most relevant studies related to Machine Learning for Water Quality Classification, the selection process was carried out in two (2) stages:

1. Applying inclusion and exclusion criteria, and

2. Conducting filtering based on the Quality Standard Questions (QSQ) of the studies.

Table 2 presents the inclusion and exclusion standards established for this SLR. All retrieved studies were screened carefully following these criteria. Research papers were included if they focused on the application of machine learning, artificial intelligence, or data-driven models in water quality monitoring, prediction, or classification, and provided at least one potential answer to the identified research questions based on the analysis of their titles, keywords, and abstracts. Studies were excluded if they were published in languages other than English, did not address water quality or machine learning, or focused on unrelated environmental or industrial domains. Finally, a duplicate analysis was performed to remove repeated entries and ensure that only the most recent and complete versions of each article were retained for review.

Table 2. Inclusion and Exclusion Standard

Inclusion Search Standard	Exclusion Search Standard
Studies must be written in the English language.	Studies that are not written in English language
Studies have potential to answer Research Questions (RQs) based on keywords, title, and abstract.	Studies will avoid duplicating the copies, review paper, and only the complete version included for this SLR.
Studies are focusing on Water Quality Classification	Studies that are not focusing on Water Quality Classification
Studies that are reporting the issues, challenges, and future enhancements of machine learning techniques for water quality classification.	Studies that are not clearly defined the concern of Water Quality Classification. Gray Studies (Study that non-publish, non-peer reviewed and work in progress)

The Inclusion and exclusion standard have been done to improve the quality of the search and produce final studies. Each of the studies were precisely studied (title, abstract and full content) and evaluated according to the Quality Standard Question (QSQ) in Table 3. The final studies will be scored as follows: Yes = 3, Moderately = 2, and No= 1. The aggregate of the answers to all the questions determines the study's overall score.

INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN SOCIAL SCIENCE (IJRISS)
ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue I January 2026

Table 3. Quality Standard Question (QSQ)

QSQ ID	Inclusion Search Standard	Exclusion Search Standard
QSQ1	Are the aims of studies clearly stated?	Yes= 3 / moderate = 2 / no = 1
QSQ2	Are the context of studies well defined?	Yes= 3 / moderate = 2 / no = 1
QSQ3	Does the study focus on RQ in the specified domain?	Yes= 3 / moderate = 2 / no = 1
QSQ4	Are the proposed classification algorithms in studies well explained?	Yes= 3 / moderate = 2 / no = 1
QSQ5	Is the proposed classification algorithm can classify the water quality pattern?	Yes= 3 / moderate = 2 / no = 1
QSQ6	Is the proposed classification algorithm compared with other classification algorithm?	Yes= 3 / moderate = 2 / no = 1
QSQ7	Is the result well explained?	Yes= 3 / moderate = 2 / no = 1

A quality score threshold of 17 was selected as it represents more than 80% of the maximum achievable score, ensuring that only studies with sufficient methodological rigor and relevance were included. Each study was independently assessed using the quality score questions. Any disagreements in scoring were resolved through discussion until consensus was achieved.

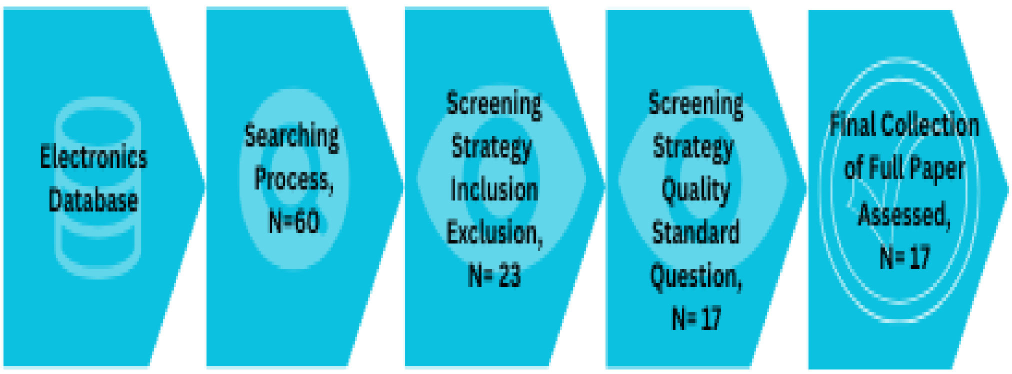

Figure 2. Final Screening Result

All the authors of the present study discussed the results and compared them to resolve all contradictions and reach a consensus. To enhance the reliability of the outcome, only the research that have quality ratings of less than 17 (Meaning that they are less than half of the maximum quality rating score of 21) will not conduct in this SLR. Figure 2 shows the result of the screening strategy. Table 4 shows the Quality Standard Question (QSQ) scores for the results of 17 studies. Figure 2 shows the screening process from three electronic databases until final collection of studies counted.

The highest-quality papers are those that achieved the maximum total QSQ score of 21, indicating full compliance with all seven quality criteria. Specifically, studies cited as [12], [13], [14], [15], [17], [18], [19], [21], [22], [25], and [28] were identified as the best papers, as each consistently scored 3 across all or nearly all QSQ items. Among these, references [12], [13], [14], [15], [17], [18], [19], [20], [21], and [22] demonstrated perfect methodological rigor with uniform maximum scores in QSQ1–QSQ7, reflecting strong research design, clarity of objectives, robust analysis, and reliable reporting. In contrast, studies [16], [23], [24], [26], and [28] obtained comparatively lower scores (17–19), indicating minor methodological limitations in specific quality dimensions. Overall, the citation-based QSQ analysis highlights [12], [13], [14], [15], [17], [22], [25], and [27] as the most reliable and methodologically sound references for this review.

RESULTS & DISCUSSION

Distribution of Studies (RQ1)

The total paper to be discussed are 17 studies from three electronic libraries which are Scopus, IEEE and Science Direct as a final screening for this SLR. These final studies consist of 17 journal papers which are from the final selected studies. Figure 3 illustrates the percentage of the publication of the final screening studies. Figure 3 shows the percentage of included studies which is IEEE is around 58.8%, Scopus is around 17.6% and Science Direct is around 23.5%.

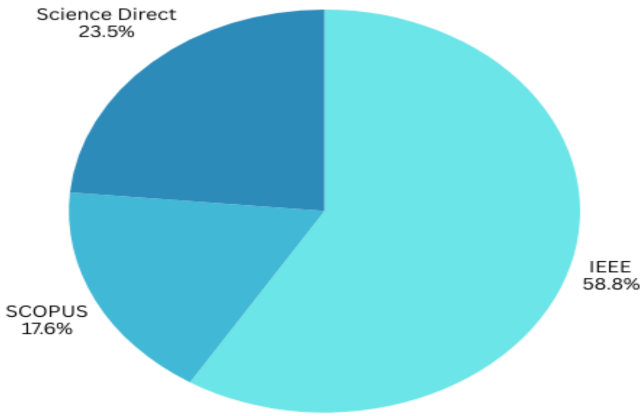

Figure 3. Percentage Included Studies

Table 4. Number of Studies after Screening Strategies

Electronic Library	Screening Process	Screening Strategy Exclusion Including		Screening Strategy Quality Standard Question (QSQ)	
IEEE	37	14	23	10	4
Science Direct	3	3	0	3	0
Scopus	20	6	14	4	2
Total	60	23	37	**17**	6

Table 4 shows the screening process of this SLR. At the initial searching process, a total of 60 studies were screened across all three electronic libraries. After going through the “Screening Strategy Exclusion and Including”, 23 studies are included and 37 studies are being excluded. Then “Screening Strategy Quality Standard Question (QSQ)” filtering the studies made only 17 studies are included and another 6 studies were excluded. Figure 4 shows the count of final collected studies over the year. The graph in Figure 4 shows the increasing slope from year 2019 until 2025.

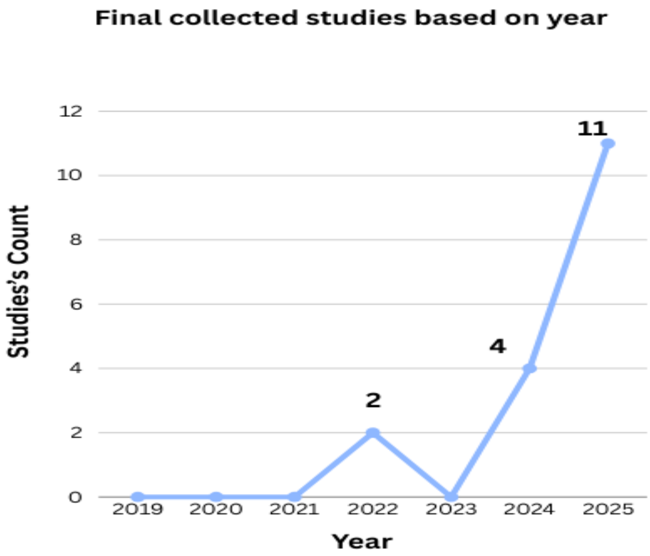

Figure 4. Count of Studies Over Year

Which are the most used algorithms are most commonly used for water quality classification?? (RQ2)

From the 17 studies reviewed, Random Forest stands out as the most used algorithm, appearing in nine (9) studies. Its popularity is no surprise Random Forest is known for being stable, reliable, and capable of handling large and diverse datasets, which are typical in water quality analysis. It also works well when different water indicators interact in complicated ways, making it a trusted tool for classifying and predicting water quality in real-world environments.

Closely following Random Forest, several other algorithms like Support Vector Machines (SVM), XGBoost, and Artificial Neural Networks (ANN) also appear frequently across the reviewed water quality studies. Their repeated use shows that researchers rely on these models for their strong performance, especially when dealing with complex and diverse water quality indicators such as pH, dissolved oxygen, turbidity, and nutrient levels. The popularity of SVM, for example, reflects its ability to handle non-linear patterns, which are common in environmental datasets where water conditions can change quickly due to weather, pollution, or human activities.

k-Nearest Neighbour (kNN) also appears in several studies, suggesting that simple and easy-to-interpret models still play an important role in water quality monitoring. kNN is especially useful when fast decision-making is needed, such as in real-time water quality alerts for drinking water or river monitoring. ANN-based methods, on the other hand, continue to gain traction because of their ability to learn deeper relationships between multiple water quality indicators, particularly when supported by good data preprocessing and feature engineering.

Deep learning approaches including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks are not as widely used yet, but their presence in the studies shows growing interest in more advanced modelling techniques. CNNs are particularly useful when working with structured or time-patterned sensor data, while LSTMs are designed to capture changes over time, making them promise for continuous water monitoring systems where conditions shift daily or even hourly. This trend indicates a gradual move toward models that can understand both spatial and temporal behaviour in water bodies.

Other noteworthy algorithms include Decision Trees, Logistic Regression, Gradient Boosting, Gaussian Process Regression, and Gating Mechanisms. These methods are often chosen for their stability, transparency, and strong performance on structured environmental datasets. XGBoost, specifically, stands out as a high-performing and efficient gradient boosting technique, valued for its accuracy and reliability even in noisy or incomplete datasets conditions common in field-based water monitoring.

Less frequently used methods such as Federated Learning, Stochastic Gradient Descent, Back-Propagation Neural Networks, and optimization-based models like Whale Optimization or Inverse Distance Weighted

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue I January 2026

methods appear only once or twice. Their limited use suggests that they are still experimental in this field, but their inclusion demonstrates ongoing exploration into new modelling strategies that might support future water quality monitoring challenges, especially as IoT sensors and distributed monitoring networks become more common.

Overall, the algorithm choices across the studies show two clear patterns: continued trust in well-established models such as Random Forest, SVM, and ANN, and a growing shift toward more sophisticated deep learning methods as datasets become larger and more complex. This diversity also highlights that there is no single "best" algorithm for water quality classification. Instead, researchers select models based on the nature of their dataset, the water indicators being measured, the availability of computational resources, and the specific goals of their monitoring system. As technology evolves, future studies are likely to combine classical machine learning with deep learning and hybrid techniques to build more accurate, scalable, and intelligent water quality prediction systems. Table 5 shows the summary of the discussion.

Table 5. Classification Algorithm for Water Quality

References	**Algorithm**	**Count**
[12], [14], [16], [18], [21], [22], [25], [26], [28]	Random Forest	9
[12], [16], [18], [21], [22], [23], [25], [28]	Support Vector Machines (SVM) Based	8
[12], [14], [16], [18], [26], [28]	XGBoost (EXtreme Gradient Boosting) algorithm	6
[14], [16], [20], [22], [25]	Artificial Neural Network (ANN) Based	5
[14], [15], [16], [18], [28]	Decision Tree (DT)	5
[13], [15], [16], [28]	k-Nearest Neighbour (kNN)	4
[15], [18], [19], [21]	Logistic Regression (LR)	4
[16], [19], [28]	CatBoost	3
[13], [16], [27]	Long Short-Term Memory (LSTM)	3
[18], [28], [29]	AdaBoost	3
[13], [16], [27]	Convolutional Neural Network (CNN) based	3
[19], [28]	Perceptron and Multilayer Perceptron (MLP) Classifier	1
[14]	Transformer-Based Model (TFM)	1
[17]	Gating Mechanism	1
[17]	Gated Liquid Neural Network (Gated-LNN)	1
[17]	Liquid Neural Network (LNN)	1
[20]	Random Tree (RT)	1

[20]	M5P	1
[20]	Reduced Error Pruning Tree (REPT)	1
[14], [19], [28]	Gradient Boosting Machine (GBM)	1
[24]	Privacy-Preserving Algorithms	1
[24]	Federated Learning (FL)	1
[25]	Gaussian Process Regression (GPR)	1
[15], [26]	Gaussian Naïve Bayes (GNB)	1
[27]	Gated Recurrent Unit (GRU)	1
[15]	Stochastic Gradient Descent Classifier (SGDC)	1
[16]	Bayesian Networks	1
[19]	Support Vector Classifier (SVC)	1
[20]	Back-Propagation Neural Network (BPNN)	2
[23]	Inverse Distance Weighted (IDW)	1
[27]	Whale Optimization Algorithm (WOA)	1

How many sensors are used, and what types are implemented for water quality classification? (RQ3)

Different types of sensors, as well as the technique by which sensor data will be utilized to create a machine learning–based water quality classification model(s) such as depth of installation, will alter both the accuracy and reliability of the model, because the type of sensors utilized will determine the system's ability to detect changes in variable water quality characteristics such as: pH, turbidity, dissolved oxygen, temperature, and nutrient levels. Therefore, knowing how many types of sensors were used in past studies is essential to assessing the overall reliability of the data used to develop a machine learning–based water quality classification model and robustness of the classification model(s), because of the different ways that sensor types (optical, electrochemical, ion-selective) respond to changing environmental conditions.

The dimensionality of the dataset also depends on how many sensors were used to gather the data. Therefore, the higher the dimensionality of the dataset, the more complex the required machine learning solution and the overall system design. To identify best practices, locate common patterns, and identify where further improvement can be made to the current water quality monitoring approach; we are looking at the types and number of sensors that have been used in water quality monitoring to determine what sensors may have been most effective for water quality classifications.

There is a great deal of variation between sensor types used in the Machine Learning Applications for Water Quality Classification research studies. A standard and proven approach to the above compilation of studies is to use 4 sensors, as observed from Zhao & O’Loughlin (2025), who combined satellite multispectral sensors (Sentinel-2 MSI, Landsat-8 OLI, and MODIS) with in-situ sensors (pH, conductivity, ORP and dissolved oxygen); and Thakkar et al. (2024), who employed in-situ analysers. The use of 4 sensors offers a good balance providing the necessary water quality parameters while still maintaining system simplicity and cost-effectiveness.

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue I January 2026

Apart from the traditional sensor configuration, various investigations have utilized specific and extensive sensor configurations to enable better characterization of the water quality. An example of this is a study by Mridha et al., 2025, who combined seven specific sensors (including SEN0161, pH; DS18B20, temperature; turbidity sensors) and a webcam for visual monitoring of the water. In another example, Aderemi et al., 2025 developed an even more comprehensive configuration with twelve different sensors measuring a wide range of parameters (pH, dissolved oxygen, turbidity, nitrates, heavy metals, Biochemical/Chemical Oxygen Demand). Azzova et al., 2025, also used six different industrial type sensors (DMA25, CAS51) for accurate chemical analysis, and Durgun 2024, used a single multispectral sensor (AS7265X) to provide spectral information for evaluation of the water.

Interestingly, many entries in the table contain no sensor information. Aslam et al. (2022) say they only used satellite images, not physical sensors; Reddy et al. (2025) mention an IoT system where many sensors exist, although the exact number of sensors was not listed. The difference between how these two studies reported what types of data were collected demonstrates a lack of transparency regarding methodology and may make it difficult to replicate findings or compare results between studies.

It appears, from the trends that have been noted so far, that researchers primarily prefer using lower-cost, compact sensor arrays for scalable monitoring. Researchers appear to prefer the four-sensor configurations for most applications. This trend holds true for most applications; however, about the need for a greater degree of resolution or an overall better, broader profile of the monitored area, especially for complicated and/or contaminated environments, researchers will very often use larger, multidimensional multi-parameter sensor arrays. Although multi-parameter sensor arrays produce richer information, they tend to lead to issues surrounding power consumption, complex calibration processes, and hardware durability, which results in them primarily being used in research scenarios instead of for routine or day-to-day operations.

It can be concluded from this literature review of all the studies using soft sensors to predict water quality, that the type and number of sensors used in ML-based studies of water quality differ significantly because of differing research goals, environmental conditions, and practical limitations. This variability highlights the necessity for researchers to clearly document sensor specifications to enable the creation of transparent, reproducible, and sustainable water quality monitoring systems. Table 6 shows the number of sensors and types.

Table 6. Number of Sensor and Type of Sensors Applied

References	Number Sensor Applied	Type of Sensor
[12]	4	Sentinel-2 MSI, Landsat-8 OLI, MODIS (on Terra), MODIS (on Aqua)
[13]	7	SEN0161, DS18B20, SEN-06617, SEN0237-A, SEN0244, TDS sensor, C525 HD Webcam
[15]	4	pH analyzer, conductivity analyzer, ORP analyzer, DO analyzer
[16]	12	pH sensors, Dissolved Oxygen (DO) sensors, Turbidity sensors, Temperature sensors, Conductivity / Total Dissolved Solids (TDS) sensors, Nitrate sensors, Chlorophyll-a sensors, Heavy metal detection sensors (e.g., for fluoride, arsenic), Chemical Oxygen Demand (COD) sensors, Biochemical Oxygen Demand (BOD) sensors, Ammonia (NH_3-N) sensors, Total Phosphorus (TP) sensors
[17]	6	Dissolved oxygen sensor, pH sensor, conductivity sensor, plus lab-based measurements for BOD, nitrate, and coliforms.
[22]	1	multispectral sensor unit (AS7265X)
[24]	6	DMA25, CAS51D, CUS52D, CPS11D, CPS12D, CLS21D

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue I January 2026

Research Question 4-7

According to Table 6, the authors present thorough evaluations of the methodologies and their corresponding metrics used to study various aspects of Water Quality. Additionally, an equivalent overview table summarises the authors' recommendations for future research in water quality classification addresses challenges associated with sensor noise, calibration issues, inconsistent sampling methods, limited cross sensitivities, and lack of standardisation in testing. Nonetheless, small imbalanced dataset issues continue to present challenges to the generalisation of model performance in practice.

Several key areas of potential growth in the classification of water quality are identified in the Future Trends column, including deep learning models that can automatically extract features from the data, improvements to high-sensitivity sensors; and integration of platform technologies using the Internet of Things (IoT) for continued, real-time water quality monitoring. The table provides an overview of the state of current research and highlights potential new technologies, each of which poses considerable technical challenges for future development before they will be operational as the next generation of water quality classification systems, accordingly, the table is also useful for answering key research questions while identifying areas of advancement in the water quality classification field.

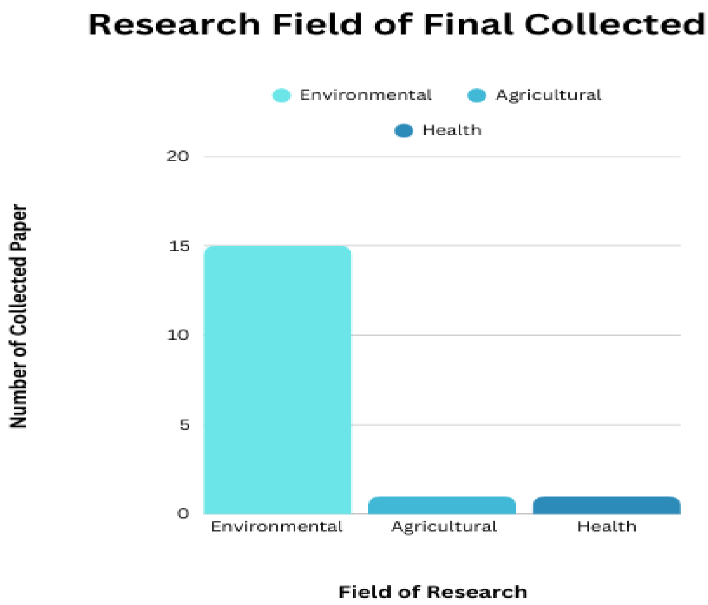

Figure 5. Research Field of Final Collected Paper

Based of Figure 5W, Environmental sensing and monitoring (ESM) are the greatest focus of the 17 studies reviewed; 15 are dedicated to the research of ESM. ESM is becoming more popular as pollution in the water ecosystem must be identified and tracked in real-time due to the increasing need to protect ecological and human health and well-being from contaminated water. Environmental sensors, such as artificial intelligence (AI) models, Internet of Things (IoT) sensors, and hybrid algorithm-based sensors, will assist researchers in realizing more accurate and faster methods of assessing the quality of water. These tools will allow researchers to conduct timely identification and monitoring of water contaminants, pathogens, and chemical shifts, thereby providing them the ability to actively manage and maintain the environment.

Agricultural sustainability (AS) is the second focus of this group of researchers; two of the studies reviewed pertain to AS. With AS, researchers emphasise smart on-farm water usage, i.e., the use of available water resources for irrigation as well as determining the best crops to grow based on water quality and potential yield. Therefore, the two research areas of ESM and AS demonstrate how researchers are using technology to maintain both human and ecological health and support an intelligent method of managing and utilizing water.

The graph shown in this report indicates a growing interest in protecting ecosystem & biodiversity (EBP) through the use of predictive-modeling techniques for assessing river and watershed health (WS/MH) and the link between that and protecting our river systems. Protecting our river systems supports two areas of human activity: 1) the need for clean drinking water for human consumption and 2) the need for healthy and balanced ecosystems through inclusive and sustainable water management practices. Thus, EBP is an example of an integrated and holistic model of water management that incorporates consideration of both social and economic factors along with environmental and ecosystem needs, thereby achieving an ecological balance. Table 7 show the summarization of Research Question from 4 until 7.

Table 7. Summarization of RQ 4-7

Ref	Metric	Limitation	Future Trend
[12]	The primary evaluation metric used is the F1-macro score.	- Single-sensor images often fail for long-term monitoring needs. - Remote sensing is limited by image acquisition time and cloud cover. - Machine learning models underestimated eutrophic cases due to limited training data.	- The multiplatform approach enhances water quality monitoring capabilities - Future work includes improving model accuracy and cross-validation across sensor.
[13]	Accuracy	- Environmental noise and lighting variability may distort image classification - Periodic model retraining may be necessary for variable effluent compositions - Non-industrial-grade sensors may affect long-term reliability in field deployments	- Enhanced image preprocessing techniques will be explored. - The sensor array will be upgraded with industrial-grade components. - Future research will deploy the system in real industrial environments.
[14]	- Accuracy - Precision - Recall - F1-Score	- The ensemble model's complexity increases interpretability challenges and resource requirements. - Model performance relies heavily on data quality and representativeness - Maintenance and updates of the model can be resource-intensive. - Applicability to other regions is limited due to training data from one state.	- Future work includes external validation with diverse water quality datasets. - The model will be tested across different Indian states and international repositories. - Integration with IoT devices for real-time data acquisition is planned.
[15]	- Accuracy - Precision - F1 score	- Water quality assessment relies on subjective analysis and limited attributes. - External factors significantly challenge water potability	- Future research will focus on advancing hybrid models for better prediction accuracy. - Integration of AI and blockchain technology is proposed for water treatment management.

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue I January 2026

		determination. - Inaccurate reporting can lead to serious health issues.	
[16]	- none	- Traditional models struggle with complex water systems and missing data. - Limited generalizability due to region-specific data and small sample sizes. - Access inequity in sensor networks affects model training and performance. - Lack of transparency in AI models complicates interpretation.	- Future research will integrate XAI with digital twins and edge computing. - Multimodal frameworks are emerging to tackle data heterogeneity
[17]	-Parameters - Ratings - Weights	- The model may struggle with extreme heterogeneity in datasets. - Errors in data pre-processing can adversely impact outcomes.	- Future work will enhance computational efficiency and generalization tests on diverse datasets. - The adaptability of the model to different regions is a key consideration.
[18]	- Accuracy - Rates - Precision -Recall	- The model's performance could improve with advanced optimization techniques. - Limitations include not using dimensionality reduction methods like PCA.	- Future studies may focus on advanced optimization techniques for better predictions. - Machine learning enhances water quality classification accuracy and efficiency. - Incorporating Deep Learning techniques like CNN can enhance quality assessment. - Ensemble models with optimized hyperparameters outperform individual models.
[19]	-Accuracy -Precision	- Incomplete temporal or spatial metadata compromises classification reliability. - MLP classifier struggles to distinguish eutrophic from noneutrophic reservoirs.	- Future studies will leverage temporal depth for detailed analyses of eutrophication variability. - The methodology allows for continuous assessment as new data becomes available.
[20]	-Accuracy -Error -Prediction -Performance	- The study used a smaller dataset over two years, limiting long-term analysis. - Statistical and ML algorithms were used; deep learning could enhance results.	- Future research can utilize long-term datasets over multiple years. - Incorporating water quality classification parameters like COD and BOD is recommended. - Deep learning algorithms could enhance

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue I January 2026

			predictive accuracy.
[21]	-Accuracy	- Combining water indices masks individual parameter effects on irrigation quality. - Sensitivity values of some networks were unsuitable for effective communication.	- Future applications may include advanced data analytics and cloud services. - The study proposes a real-time water monitoring network using Machine Learning tools.
[22]	-Accuracy -Precision -Recall -F1 score	- Challenges exist in applying techniques to varied water types and conditions. - Long-term studies are required to assess reliability and consistency. - Broader validation across different environments is needed for these methods.	- Advanced machine learning algorithms will enhance water quality monitoring systems. - Integration of AI and spectral analysis is a promising future trend. - High-resolution remote sensing data will play a significant role in water quality assessment.
[24]	-Accuracy -Precision -Recall -F1 score	- High complexity in applying differential privacy requires further auditing methods. - Heterogeneous data in federated learning presents unresolved challenges for differential privacy.	- Differential privacy can enhance water quality monitoring accuracy with minimal performance loss. - Future research may explore more neural network architectures with DP applications. - The study indicates potential for improved anomaly detection in water monitoring systems.
[23]	-Error -Precision -Accuracy	- Unequal distribution of water sample types affects accuracy. - Limited data restricts widespread application of water quality classification models. - Historical data limits model training effectiveness.	- The study recommends combining Entropy, TOPSIS, SMOTE, and SVM methodologies for forecasting.
[25]	-Accuracy -Error -Precision -F1 score -Various	- Machine learning faces challenges with data quality and model interpretability. - Scalability issues require careful consideration in real-world applications. - Complex relationships in water quality factors complicate predictions.	- Future forecasting of water quality is possible with machine learning advancements.

[26]	-Accuracy -Precision -Recall -F1 score	- Regional data variability introduces uncertainty in predictions. - Current recommendations rely on static mappings, limiting decision precision. - IoT sensor noise can compromise prediction reliability.	- The framework supports rapid decision-making for farmers and agricultural agencies. - It utilizes machine learning for real-time predictions and crop recommendations.
[27]	-Accuracy -Recall -F1 score -Presicion	- Traditional methods face limitations like single-model approaches and inadequate data processing capabilities. - These limitations hinder effective water quality monitoring and prediction accuracy.	- Future research will focus on data accuracy and veracity. - The study aims to enhance water quality prediction model performance.
[28]	-Accuracy -Precision -Recall -F1 score -Model	- Data availability and computational complexity hinder machine learning applications. - Model interpretability remains a significant challenge for widespread adoption. - Lack of standardized evaluation metrics affects model generalization.	- Future innovations include explainable AI and AutoML for improved predictions.

CONCLUSION

This systematic literature review provides an overview of sensor technologies and classification algorithms for assessing water quality. The Random Forest, Support Vector Machine (SVM), (ANN), and k-Nearest Neighbor (kNN) are the models most used because they can process complicated and nonlinear data associated with water quality. Additionally, there are several examples of multi-parameter sensors that measure multiple parameters such as pH, turbidity, Dissolved Oxygen (DO), and EC; these devices offer a good trade-off between the cost, input and output accuracy, and ease of deployment. Although advances have been made in this area, many challenges still exist, such as sensor calibration issues, inconsistent evaluation methods, and limited application of the models across various water sources. Future studies should explore AI-based adaptive learning systems, compact and low power sensor designs, and innovative data fusion methods to improve reliability and provide real-time decision-support. In addition, establishing a standardized benchmarking framework and promoting collaboration between disciplines will be essential in increasing the scalability and utilization of these technologies. Thus, this review provides the basis for improving the classification of water quality and smart monitoring systems.

ACKNOWLEDGMENTS

The authors express deepest gratitude to Universiti Teknikal Malaysia Melaka (UTeM) for their unwavering support and encouragement throughout this research journey under Grant PJP/2024/FTMK/PERINTIS/SA0036. Their facilities, resources, and academic environment have been instrumental in the successful completion of this work. Also, to Universiti Malaysia Pahang Al-Sultan Abdullah and Universitas Amikom for the contribution in this article.

REFERENCES

1. A. Lokman, W. Z. W. Ismail, and N. A. A. Aziz, "A Review of Water Quality Forecasting and Classification Using Machine Learning Models and Statistical Analysis," Aug. 01, 2025, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/w17152243.
2. W. Chen, D. Xu, B. Pan, Y. Zhao, and Y. Song, "Machine Learning-Based Water Quality Classification Assessment," Water (Switzerland), vol. 16, no. 20, Oct. 2024, doi: 10.3390/w16202951.
3. H. Yang, J. Kong, H. Hu, Y. Du, M. Gao, and F. Chen, "A Review of Remote Sensing for Water Quality Retrieval: Progress and Challenges," Remote Sens (Basel), vol. 14, no. 8, Apr. 2022, doi: 10.3390/rs14081770.
4. P. Talukdar, B. Kumar, and V. V. Kulkarni, "A review of water quality models and monitoring methods for capabilities of pollutant source identification, classification, and transport simulation," Sep. 01, 2023, Springer Science and Business Media B.V. doi: 10.1007/s11157-023-09658-z.
5. M. T. Ejigu, "Overview of water quality modeling," Cogent Eng, vol. 8, no. 1, 2021, doi: 10.1080/23311916.2021.1891711.
6. T. Yan, S. L. Shen, and A. Zhou, "Indices and models of surface water quality assessment: Review and perspectives," Sep. 01, 2022, Elsevier Ltd. doi: 10.1016/j.envpol.2022.119611.
7. B. Kitchenham et al., "Systematic literature reviews in software engineering-A tertiary study," 2010, Elsevier B.V. doi: 10.1016/j.infsof.2010.03.006.
8. A. F. J. AL-Gburi, M. Z. A. Nazri, M. R. Bin Yaakub, and Z. A. A. Alyasseri, "A systematic review of symbiotic organisms search algorithm for data clustering and predictive analysis," Jan. 01, 2024, Walter de Gruyter GmbH. doi: 10.1515/jisys-2023-0267.
9. B. Hammouti et al., "Bibliometric analysis of global research trends on UMI using Scopus database and VOS viewer from 1987-2024," J. Mater. Environ. Sci, vol. 2025, no. 4, pp. 548–561, 2025, [Online]. Available: http://www.jmaterenvironsci.comhttp://www.jmaterenvironsci.com
10. M. Kumar and S. A. Librarian, "Scientific Research on Cutting-Edge Technology : a scientometric approach on IEEE Xplore Digital Library."
11. A. Capari, H. Azarbonyad, G. Tsatsaronis, Z. Afzal, and J. Dunham, "ScienceDirect Topic Pages: A Knowledge Base of Scientific Concepts Across Various Science Domains," in SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, Inc, Jul. 2024, pp. 2976–2980. doi: 10.1145/3626772.3661353.
12. M. Zhao and F. O'Loughlin, "A Multiplatform Approach for Chlorophyll Level Estimation for Irish Lakes," IEEE J Sel Top Appl Earth Obs Remote Sens, vol. 18, pp. 8261–8274, 2025, doi: 10.1109/JSTARS.2025.3546060.
13. M. J. H. Mridha et al., "A Real-Time ETP Outlet Monitoring Framework Leveraging Environmental IoT, Colorimetry, and Learning Theory," IEEE Access, vol. 13, pp. 98729–98746, 2025, doi: 10.1109/ACCESS.2025.3576826.
14. J. K. Pandya, S. S. Khandelwal, R. K. Tipu, and K. S. Pandya, "Advancing Water Quality Management: An Integrated Approach Using Ensemble Machine Learning and Real-Time Interactive Visualization," IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3573589.
15. D. S. Thakkar et al., "Blockchain-Orchestrated Intelligent Water Treatment Plant Profiling Framework to Enhance Human Life Expectancy," IEEE Access, vol. 12, pp. 49151–49166, 2024, doi: 10.1109/ACCESS.2024.3384607.
16. I. A. Aderemi, T. O. Kehinde, U. Daniel Okwor, K. H. Ahmad, K. Y. Adjei, and C. Cyriacus Ekechi, "Explainable AI for Water Quality Monitoring: A Systematic Review of Transparency, Interpretability, and Trust," IEEE Sensors Reviews, vol. 2, pp. 419–443, Aug. 2025, doi: 10.1109/sr.2025.3595500.
17. S. Chadalavada et al., "Gated-LNN: Gated Liquid Neural Networks for Accurate Water Quality Index Prediction and Classification," IEEE Access, vol. 13, pp. 69500–69512, 2025, doi: 10.1109/ACCESS.2025.3561593.
18. Z. Karami Lawal et al., "Optimized Ensemble Methods for Classifying Imbalanced Water Quality Index Data," IEEE Access, vol. 12, pp. 178536–178551, 2024, doi: 10.1109/ACCESS.2024.3502361.
19. R. Usamentiaga, J. Sal, and P. Elvira, "Remote Sensing and Machine Learning for Eutrophication Detection: Assessing the Trophic State in Reservoirs Using Multispectral Indices and Deep Learning,"

IEEE J Sel Top Appl Earth Obs Remote Sens, vol. 18, pp. 16206–16226, 2025, doi: 10.1109/JSTARS.2025.3583761.

20. B. Aslam, A. Maqsoom, A. H. Cheema, F. Ullah, A. Alharbi, and M. Imran, “Water Quality Management Using Hybrid Machine Learning and Data Mining Algorithms: An Indexing Approach,” IEEE Access, vol. 10, pp. 119692–119705, 2022, doi: 10.1109/ACCESS.2022.3221430.
21. O. O. Ajayi, A. B. Bagula, H. C. Maluleke, Z. Gaffoor, N. Jovanovic, and K. C. Pietersen, “WaterNet: A Network for Monitoring and Assessing Water Quality for Drinking and Irrigation Purposes,” IEEE Access, vol. 10, pp. 48318–48337, 2022, doi: 10.1109/ACCESS.2022.3172274.
22. Y. Durgun, “Real-time water quality monitoring using AI-enabled sensors: Detection of contaminants and UV disinfection analysis in smart urban water systems,” J King Saud Univ Sci, vol. 36, no. 9, Oct. 2024, doi: 10.1016/j.jksus.2024.103409.
23. A. Das, “Water quality assessment and geospatial techniques for the delineation of surface water potential zones: A data-driven approach using machine learning models,” Desalination Water Treat, vol. 324, Oct. 2025, doi: 10.1016/j.dwt.2025.101461.
24. A. Arzovs, S. Parshutin, V. Urbanovics, J. Rubulis, and S. Dejus, “Application of differential privacy to sensor data in water quality monitoring task,” Ecol Inform, vol. 86, May 2025, doi: 10.1016/j.ecoinf.2025.103019.
25. D. Irwan et al., “River water quality monitoring using machine learning with multiple possible in-situ scenarios,” Environmental and Sustainability Indicators, vol. 26, Jun. 2025, doi: 10.1016/j.indic.2025.100620.
26. C. Reddy, V. V. Reddy, J. Chetan Vikas, and K. Priyadarsini, “AI-Driven Water Quality Assessment and Crop Suitability System,” in 2025 International Conference on Data Science and Business Systems, ICDSBS 2025, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/ICDSBS63635.2025.11031738.
27. M. Wu and E. B. Blancaflor, “Research on Watershed Water Quality Classification Prediction Based on WOA-CNN-GRU Model,” in E3S Web of Conferences, EDP Sciences, Oct. 2024. doi: 10.1051/e3sconf/202458001007.
28. N. Arunachalam, S. Nilakhe, and V. Dhing, “Water Quality Prediction Using Machine Learning Models,” in Proceedings of the 2025 11th International Conference on Communication and Signal Processing, ICCSP 2025, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 53–58. doi: 10.1109/ICCSP64183.2025.11088772.
29. C. Cheng et al., “Urban Fine-Grained Water Quality Monitoring Based on Stacked Machine Learning Approach,” IEEE Access, vol. 12, pp. 77156–77170, 2024, doi: 10.1109/ACCESS.2024.3404068.