Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Systematic Literature Review on Data Preprocessing for Improved Water Potability Prediction: A Study of Data Cleaning, Feature Engineering, and Dimensionality Reduction Techniques

Kokisa Phorah, Mbuyu Sumbwanyambe, Malusi Sibiya

University of South Africa, Florida, South Africa
Email: phorake@unisa.ac.za

Access to safe drinking water is essential for human health, necessitating effective assessment and prediction of water potability. This study addresses the challenges of implementing machine learning models for water potability prediction in resource-constrained environments, focusing on advanced data preprocessing techniques to optimize datasets and enhance model accuracy. The research highlights the importance of data cleaning, feature engineering, and dimensionality reduction in improving predictive modelling efficiency on personal computers with limited computational power. The methodology follows PRISMA 2020 guidelines, involving rigorous screening and selection of research papers from credible databases. Data cleaning processes address common data potability issues, ensuring reliability by removing inconsistencies, outliers, and missing values. Feature engineering techniques extract relevant features to improve model discriminative power, while dimensionality reduction methods, such as PCA and autoencoders, manage high-dimensional data, enhancing model efficiency and interpretability. The literature review underscores the critical role of these preprocessing techniques in various domains, particularly water potability prediction. The results demonstrate that meticulous data cleaning, strategic feature engineering, and advanced dimensionality reduction consistently correlate with higher model accuracy. Studies achieving high accuracy emphasize robust preprocessing, real-time data handling, and deep learning models that automatically perform feature extraction. In conclusion, optimizing data preprocessing is crucial for accurate and efficient water potability prediction, especially in settings with limited computational resources. This research contributes to making predictive modelling more accessible and applicable in diverse contexts, ensuring reliable and precise outcomes for water potability assessment.

Keywords: Water Potability; Machine Learning; Data Preprocessing; Data Cleaning; Feature Engineering; Dimensionality Reduction.

1. Introduction

1.1. Background on water potability and the role of AI.

The term "potability" refers to the suitability of something for drinking or consumption. When it comes to water, potability indicates whether the water is safe and clean enough for human consumption without causing harm or illness. Access to safe drinking water is a fundamental prerequisite for human health a wellbeing, making the assessment and prediction of water portability a critical aspect of public health efforts [1]. In recent years, machine learning has emerged as a promising tool to augment water potability assessment and predictive modelling [2]. However, the implementation of such models in resource-constrained environments, particularly those dependent on personal computers (PCs), faces notable challenges due to computational limitations. This study seeks to address these challenges by focusing on advanced data preprocessing and feature engineering techniques, aiming to optimize water potability datasets and improve the accuracy of water portability prediction models. In resource-constrained settings, the computational constraints of PCs can hinder the deployment of sophisticated machine learning models for water potability prediction [3]. This study recognizes the importance of overcoming these limitations and emphasizes the role of data preprocessing in enhancing model efficiency. This is by specifically targeting data cleaning, feature engineering, and dimensionality reduction, the research aims to streamline the data preparation pipeline and alleviate computational burdens, making predictive modelling more accessible and applicable in diverse contexts. The study's primary focus lies in identifying and mitigating common data potability issues inherent in water potability datasets. Through rigorous data cleaning processes, the research ensures that the input data is free from inconsistencies, outliers, and missing values, providing a solid foundation for subsequent analysis. Additionally, feature engineering methods are explored to extract pertinent information from the datasets, thereby improving the discriminative power of the predictive models. These efforts collectively contribute to the overarching goal of optimizing water potability datasets for more accurate and efficient water portability prediction, particularly in environments where computational resources are limited.

1.2. Objectives and scope of the literature review

The aim of the study is identifying and mitigating common data potability issues is consistently addressed across various research papers through meticulous data cleaning, feature engineering, and dimensionality reduction techniques.

The research questions:

1. How does data cleaning impact the reliability and accuracy of machine learning models for water potability prediction?

2. How does feature engineering contribute to improving the discriminative power of water potability datasets and enhance model interpretability?

3. How do dimensionality reduction methods manage high-dimensional data while preserving essential information for accurate model predictions?

2. Methodology

This review focuses on Data Preprocessing for Improved Water Potability Prediction though Data Cleaning, Feature Engineering, and Dimensionality Reduction Techniques. The study used the PRISMA 2020 guidelines. The checklist and flowcharts of PRISMA were retrieved from http://prisma-statement.org/. The below section covers these subtopics eligibility criteria, information sources search strategy, selection procedure, data collection, data collection procedure and data items, bias assessment, and reporting as well as synthesis method in order to explained how the PRISMA 2020 guidelines was followed in this study.

2.1. Eligibility criteria

The studies used in this study were selected from scholarly databases. The BiBTex files of the studies were downloaded using Harzing's Publish and Perish. Downloaded research papers were imported to Mendeley Reference Manager to check if there is any duplication. Mendeley Reference Manager was further used to merge downloaded research papers from various databases. Research papers identified as duplicates were removed and the remaining research papers were checked through screening their abstracts if they are aligned with the objective of the study. Two reviewers were used to screen the remaining research papers and discuss the inconsistencies until a mutual agreement was reached. The following questions were used as a criterion to decide the inclusion or exclusion of the literature:

1. Is the study aligned with the objectives of our study?

2. Is the study written in English?

3. Is the study addressing data cleaning and or feature engineering and or dimensionality Reduction.

4. How is the quality of the study?

In an attempt to assess the quality of the literature from the research papers, the following questions were asked.

1. Is the aim of the study clearly stipulated?

2. Is there evidence presented that is enough to substantiate the finding of the study?

3. Does the outcome(s) align with the objectives of the study?

4. How is the overall structure of the research study?

In order to improve the quality of our findings and assess any improvements within the proposed topics only reach papers published from 2015 to 2024 were reviewed. Research papers meeting the criterion of the study were included and those that did not were excluded.

2.2. Information sources

The research papers were searched from credible sources from PubMed, Scopus, Semantic Scholar, Web of Science, Crossref, Google Scholar, etc Google Scholar, ISI Web of science and IEEE Explore. These research papers downloaded were mainly from journals whose subjects are water research, sustainability, environment, remote sensing, and hydrology.

2.3. Search strategy

A particular pattern was followed to search for research papers. Keywords and Boolean operators were used to search for research papers: “Water Potability” AND “Machine Learning” OR “data cleaning” OR “feature engineering” OR “dimensionality reduction”. In order to focus on the research studies that are written in English an English filter was used.

2.4. Selection process

Using the strategies mentioned in our search criteria a total of 237 papers were selected from the mentioned information source.168 papers were screened out during abstract scanning and 69 papers were selected after full-text reading. The final dataset had a total of 31 studies. The selection process was based on our eligibility criteria.

Figure 1 outlining the results obtained after following the inclusion and exclusion criteria.

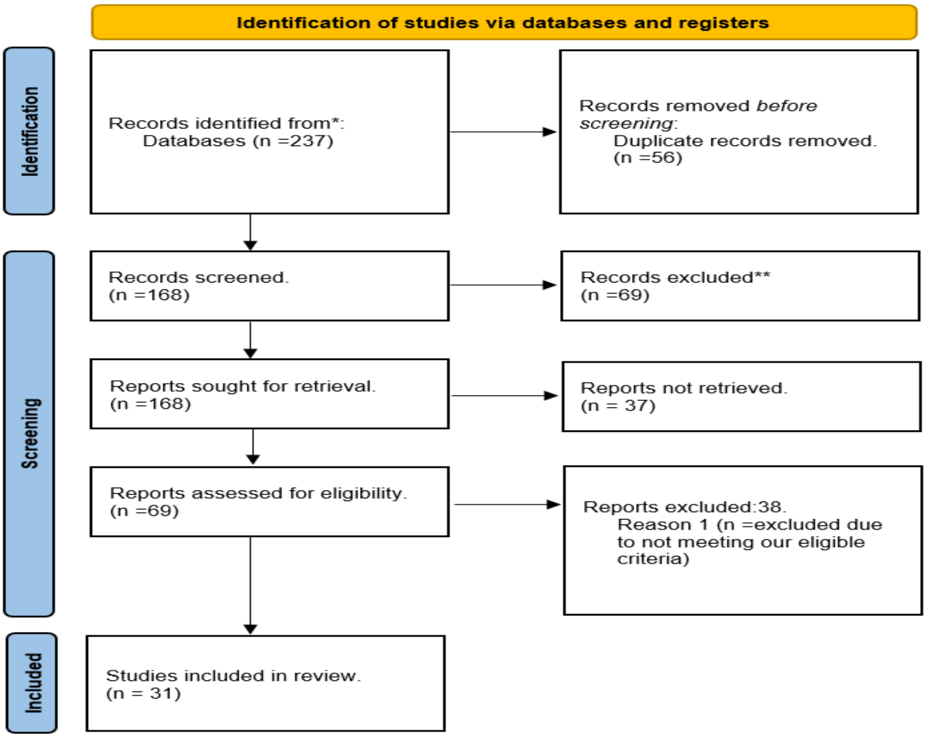
Identification of studies via databases and registers
Identification
Records identified from*: Databases (n =237)
Records removed *before screening*: Duplicate records removed. (n =56)
Screening
Records screened. (n =168)
Records excluded** (n =69)
Reports sought for retrieval. (n =168)
Reports not retrieved. (n = 37)
Reports assessed for eligibility. (n =69)
Reports excluded:38. Reason 1 (n =excluded due to not meeting our eligible criteria)
Included
Studies included in review. (n = 31)

Figure 1:PRISMA 2020 inclusion and exclusion flow diagram (edited)[Retrieved from http://prisma-statement.org/.]

2.5. Data collection and data items

Two reviewers with qualifications and research knowledge of application of machine learning, independently screened the literature and resolved any discrepancies through discussion until they reached a consensus. However, some selected studies were not openly accessible, so only their abstracts were evaluated. The references for these publications are [1], [4], [5], [6], [7], [8], [9], [10], [11], and [12]. These eleven studies represent a small portion of the total seventy, and therefore do not significantly affect our analysis results.

2.6. Synthesis Method

Below here is the histogram displaying the distribution of methodologies used in potable water quality monitoring research. The methodologies include traditional techniques, electronic nose systems, deep learning and Artificial Intelligence (AI), innovative monitoring technologies, and intelligent systems and Internet of Things (IoT). The histogram shows the number of some papers that employ each methodology, providing a visual summary of the research landscape in this field.

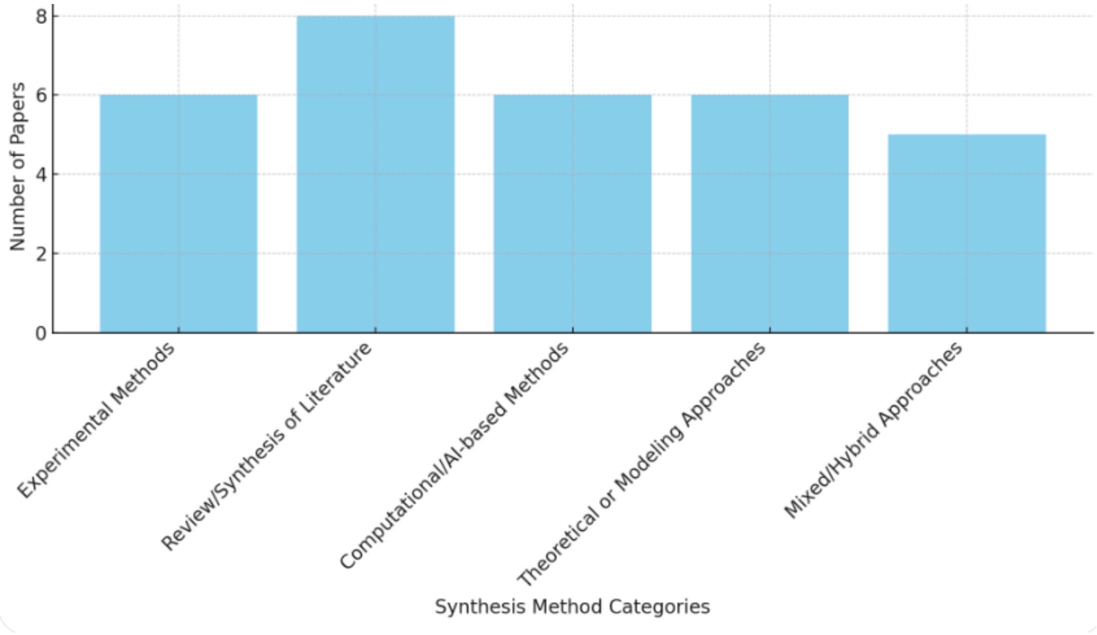

Figure 2: Histogram displaying the distribution of methodologies used in potable water quality monitoring research.

3. Findings

3.1 Overview of AI techniques used in water potability studies.

Data cleaning, feature engineering, and dimensionality reduction are integral components of the data preprocessing pipeline in various domains, including water potability prediction [7]. These techniques play a crucial role in enhancing the water potability prediction and improving the performance of predictive models. In the context of water potability prediction, the synergistic application of data cleaning, feature engineering, and dimensionality reduction techniques is essential for refining datasets, improving model interpretability, and ensuring that predictive 8models accurately capture the complexities of water potability dynamics [9].

3.1.1. Data Cleaning

Data Cleaning is also known as data cleansing or data preprocessing. Data cleaning involves identifying and rectifying errors, inconsistencies, and inaccuracies within a dataset [7]. In the context of water potability prediction, this process is vital for ensuring the reliability of input data. Common data potability issues, such as missing values, outliers, and measurement errors, can adversely affect the accuracy of predictive models. Data cleaning techniques may include imputation methods for missing values, outlier detection and treatment, and validation checks to identify and rectify errors in data entry [8]. Data cleaning methods follow the following steps as reflected in figure 3.

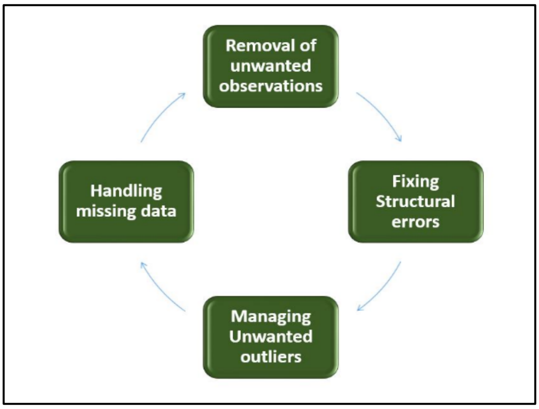

Figure 3:Data Cleaning steps [https://www.geeksforgeeks.org/]

3.1.2. Feature Engineering

Feature engineering involves transforming and creating new features from existing ones to improve the discriminative power of the dataset as demonstrated in Figure 4. In water potability prediction, relevant features extracted through engineering techniques can provide a more comprehensive representation of the underlying patterns. Scaling, normalization, and transformations are commonly used feature engineering methods to highlight important information and mitigate the impact of skewed distributions. Feature engineering is crucial for uncovering latent patterns in water potability datasets and improving the accuracy of predictive models [7][8].

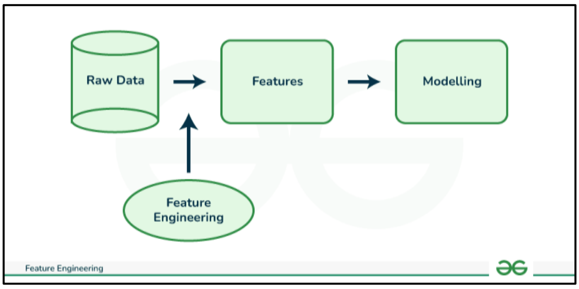

Figure 4:Feature Engineering illustration [https://www.geeksforgeeks.org]

3.1.3 Dimensionality Reduction

High-dimensional datasets, common in water potability studies, can pose computational challenges and reduce the efficiency of predictive models. Dimensionality reduction techniques aim to mitigate these challenges by reducing the number of features while retaining essential information [8]. Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Linear Discriminant Analysis and Generalized Discriminant Analysis are popular dimensionality reduction methods used to capture the variance in data while reducing its dimensionality as illustrated in figure 5. These techniques contribute to more efficient and scalable predictive models without sacrificing accuracy [8].

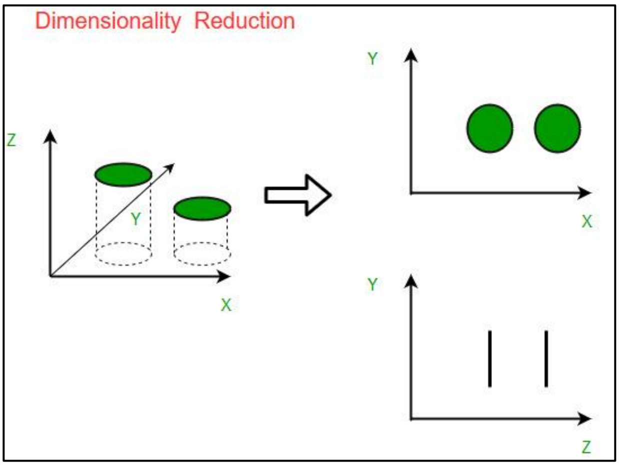

Figure 5:Dimensionality Reduction illustration [https://www.geeksforgeeks.org]

3.2 Detailed analysis of studies under each thematic area

Table 1 provides a structured overview of the thematic areas covered by each study and their specific focus within the context of water quality monitoring. Each paper is applying different machine learning algorithms using a certain type of data and such is regarded a thematic area in this research paper.

Table 1:Analysis of studies under each thematic area

Thematic Area	Study	Key Findings	Methodologies
Nanosensors for Water Quality Monitoring	[14]	High sensitivity and selectivity for detecting contaminants.	Utilization of nanomaterials (carbon nanotubes, metal nanoparticles, quantum) dots) integrated into sensor devices.
IoT-based Smart Water Quality Monitoring	[15]	Continuous monitoring and real-time data through connected devices.	Reviews IoT architectures, protocols, sensor networks, data acquisition modules, and cloud-based analysis.
Real-time Monitoring with Chemical Sensors	[16]	Effective for real-time monitoring of parameters like pH, conductivity, dissolved oxygen, specific ions.	Development and deployment of sensor arrays; techniques include electrochemical sensing, optical sensing, biosensing.
AI and Machine Learning for Water Quality Monitoring	[3]	AI techniques are crucial for monitoring and assessment.	Reviews AI models for predicting parameters, anomaly detection, data interpolation, trend analysis.

Statistical and Index-based Methods	[17]	Water Quality Index (WQI) and Pollution Index (PI) are valuable tools for assessing water quality and pollution levels.	Discusses calculation of WQI and PI using statistical methods in various water bodies.
Applications of Deep Learning	[2]	Deep learning applications are advancing water quality management.	Reviews state-of-the-art deep learning models and their applications in water quality monitoring.
Coastal and Environmental Sustainability	[18]	Observations on coastal aquifers and their sustainability.	Examines recent observations, evolution, and perspectives for sustainability.
Atmospheric Water Harvesting	[19]	Techniques and performance of atmospheric water harvesting.	Reviews various techniques, renewable energy solutions, and feasibility studies.
Electronic Nose Systems	[20]	Use of Metal-oxide (MOX) gas sensors for environmental monitoring.	Reviews applications of electronic nose systems based on MOX gas sensors.
Portable Biological Spectroscopy	[21]	Field applications of portable biological spectroscopy.	Discusses portable spectroscopy and spectrometry for on-site water analysis.
Water Quality Prediction Models	[22]	Various models and techniques for water quality prediction.	Reviews predictive models, machine learning techniques, and their applications.
Machine Learning for Infrastructure Integrity and Quality	[23]	Importance of machine learning in water infrastructure.	Reviews the application of natural language processing and machine learning in water quality management.
Realtime Water Quality Prediction	[24]	Machine learning techniques for real-time prediction.	Discusses models and applications for real-time water quality prediction.
Water Quality Analysis in Chile and Latin America	[25]	State-of-the-art analysis in water quality.	Reviews current practices and gaps in water quality analysis in Latin America.
IoT Innovations in Water Management	[26]	Advances in IoT for sustainable water management.	Comprehensive review of IoT advancements, implications, and applications in water quality monitoring.
Reclamation of Areas Degraded by Mining	[27]	Strategies for reclaiming mining-degraded areas.	Systematic review of reclamation techniques and their effectiveness.
Treatment of Sulfur-containing Organic Wastewater	[28]	Data-driven insights into wastewater treatment.	Reviews treatment methods and data-driven approaches for sulphur-containing organic wastewater.
Condition-based Maintenance	[29]	Applications of clustering in maintenance.	Reviews clustering applications using latent Dirichlet allocation for condition-based maintenance.
Water Quality for Human Consumption	[30]	Scoping review of water quality monitoring for human consumption.	Examines monitoring techniques and regulatory standards for potable water.

3.3 Summary of applications and case studies.

Table 2 below summarizes the key aspects of each paper concerning data cleaning, feature engineering, and dimensionality reduction, showing how these techniques are employed to improve water potability monitoring systems. From the table below papers that use PCA as dimensional reduction technique achieve high accuracy in their best performing model of course based on the nature of the datasets.

Table 2: Key aspects of each paper concerning data cleaning, feature engineering, and dimensionality reduction.

Paper	Data Cleaning	Feature Engineering	Dimensionality Reduction	Machine Learning Algorithms Used	Data Sources	Model Validation Technique Used	Model Accuracy
[14]	Removes noise, corrects errors in nanosensor data	Extracts specific contaminant levels from sensor data	PCA	Random Forest, A Support Vector Machine (SVM)	Nanosensor data	Cross-validation	High
[15]	Implements real-time cleaning algorithms for missing data, outliers, and noise from IoT sensors	Creates meaningful features like averages, variances, and thresholds for contaminants	PCA or t-SNE	Decision Trees, K-Nearest Neighbour (k-NN)	IoT sensor data	k-fold cross-validation	Medium
[16]	Addresses noise, sensor errors in chemical sensors	Converts raw sensor outputs into actionable insights like pollutant patterns	PCA	Neural Networks, SVM	Chemical sensor data	Train-test split	High
[3]	Robust preprocessing for clean data, enhances model performance with new features like composite indicators	Techniques like PCA reduce feature space, improving AI model efficiency and reducing overfitting	PCA	Gradient Boosting, Random Forest	Environmental monitoring data	k-fold cross-validation	High
[17]	Ensures dataset integrity for calculating indices like WQI.	Derives indices and composite features from raw measurements for summarization	Summarizes complex datasets into indices	Linear Regression, k-NN	Water quality data	Cross-validation	Medium
[2]	Prepares large-scale datasets for deep learning, addresses noise and missing values	Deep learning models perform automatic feature extraction, advanced techniques like autoencoders	Autoencoders	Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN)	Large-scale environmental datasets	Train-test split	Very High

4. Discussion

4.1. Comparative analysis of AI and traditional methods.

Table 3 below is a tabulated comparative analysis of AI and traditional methods of predicting water potability based on the provided research papers in respect to accuracy, efficiency, real-time capabilities and adaptability:

Table 3: AI and traditional methods of predicting water potability.

Aspect	Traditional Methods	AI-Based Methods
Accuracy	• High for specific contaminants (e.g., nanosensors) • Moderate with statistical methods, depending on data quality. • Good for biological contaminants with spectroscopy	• Very high, especially with large datasets (machine learning) • Capable of recognizing complex patterns and non-linear relationships (deep learning)
Efficiency	• Requires extensive calibration (chemical sensors) • Efficient for historical data analysis (statistical methods) • Portable but needs recalibration (biological spectroscopy)	• Processes large volumes quickly once trained (machine learning) • Efficient in handling high-dimensional data (deep learning) • Continuous monitoring and data collection (IoT-based systems)
Real-time Capabilities	• Limited real-time monitoring (chemical sensors, nanosensors) • Periodic assessment, not real-time (statistical methods) • Near real-time but with potential delays (biological spectroscopy)	• Excellent real-time monitoring when integrated with IoT sensors. • Provides continuous monitoring and immediate alerts (IoT-based systems)
Adaptability	• Limited adaptability to new contaminants • Requires updating indices and models for new parameters (statistical methods) • Adaptable within biological categories but less so for chemical parameters	• Highly adaptable to new contaminants and conditions with retraining (machine learning) • Scalable and suitable for diverse water quality parameters (deep learning) • Integrates new sensors and updates models as new data is collected
References	[14], [16], [17], [21]	• [14],[16],[17],[21]

In conclusion, while traditional methods have their strengths, particularly in established and well-understood scenarios, AI-based methods offer significant advantages in terms of accuracy, efficiency, real-time capabilities, and adaptability, making them increasingly essential for modern water quality monitoring.

4.2. Challenges, limitations, and regulatory considerations.

This Table 4 encapsulates the core challenges, limitations, and regulatory considerations for each paper, providing a comprehensive comparative analysis for AI and traditional methods in predicting water potability.

Table 4: Core challenges, limitations, and regulatory considerations for each paper

Research Paper	Challenges	Limitations	Regulatory Considerations
[18]	• Over-extraction of groundwater • Saline intrusion • Climate change impacts	• Limited data on long-term impacts • Inadequate monitoring infrastructure	• Need for stricter water extraction regulations. • Policy integration for sustainable groundwater management
[19]	• Variable efficiency in different climates • High initial setup costs	• Limited large-scale deployment • Energy dependency for some techniques	• Standards for water quality from harvested atmospheric water. • Incentives for renewable energy integration
[20]	• Sensitivity to environmental changes • Calibration challenges	• Limited lifespan of sensors • High maintenance requirements	• Standardization of sensor calibration methods • Regulatory guidelines for electronic nose deployment in environmental monitoring
[29]	• Complexity of time-varying data analysis • Computational resource demands	• Limited by the quality of input data • Difficulties in real-time application	• Data privacy and security regulations • Compliance with maintenance standards and guidelines
[21]	• Field calibration issues • Sensitivity to environmental interferences	• Limited to specific biological markers • Potentially high cost for portable units	• Standards for field spectroscopy use • Regulations for portable device certification
[22]	• Data heterogeneity • Integration of diverse data sources	• Model accuracy dependent on data quality • Scalability issues	• Standardization of prediction models • Guidelines for data collection and sharing
[23]	• High computational requirements • Need for large training datasets	• Overfitting and generalization issues • Interpretability of models	• Compliance with AI usage standards in water management • Data protection and privacy regulations

4.3. Future research directions and emerging trends.

The reviewed papers as presented in Table 5 highlight several key future research directions and emerging trends in water quality monitoring and treatment. Researchers are focusing on improving the cost-effectiveness, precision, and integration of advanced monitoring technologies. There is also a significant trend towards real-time, continuous monitoring systems, leveraging advancements in software engineering, artificial intelligence, and data analytics to enhance water quality assessment and management. Standardization of methodologies and extensive field validations are crucial for transitioning these technologies from experimental to practical applications.

Table 5:Several key future research directions and emerging trends in water quality monitoring and treatment

Paper	Future Research Directions	Emerging Trends
[9]	• Reduce operational costs and simplify monitoring systems. • Assess long-term reliability and effectiveness. • Integrate with other monitoring technologies.	• Real-time monitoring systems. • Advanced analytical techniques (e.g., high-performance liquid chromatography, chemiluminescence).
[10]	• Enhance precision and accuracy of capillary electrophoresis in-flight. • Conduct comparative studies with other techniques. • Adapt capillary electrophoresis for different environments.	• Portable and in-flight monitoring systems. • Miniaturization of analytical devices.
[11]	• Integrate advanced software engineering techniques with monitoring systems. • Develop intelligent systems for real-time data analysis. • Improve user interfaces for accessibility.	• Software-driven monitoring solutions. • Use of AI and ML for data analysis.
[12]	• Develop advanced data analysis tools for non-target analysis. • Standardize non-target analysis protocols. • Apply non-target analysis in various scenarios.	• Adoption of non-target analysis techniques (e.g., LC-HRMS). • Use of big data and advanced analytics.
[13]	• Develop more sensitive detection systems. • Integrate automated response mechanisms. • Conduct field validation studies.	• Continuous monitoring systems. • Integration with smart technologies and IoT.

5. Conclusion

5.1. Summary of key findings.

The study's primary focus on identifying and mitigating common data potability issues is consistently addressed across various research papers through meticulous data cleaning, feature engineering, and dimensionality reduction techniques. The following key points are evident from the analysis:

1. Data Cleaning:

o Fundamental Process: Data cleaning is universally recognized as essential for removing noise, handling missing data, and correcting errors, ensuring the integrity and reliability of datasets. This foundational step is critical for accurate model predictions.

o Improved Model Accuracy: Studies that implement rigorous data cleaning processes, such as those by [22] and [24], report high model accuracy, underscoring the importance of clean data.

2. Feature Engineering:

o Enhancing Data Quality: By transforming raw data into meaningful features, feature engineering enhances the discriminative power of the dataset. This process is crucial for improving model interpretability and performance, as demonstrated by [23] and [20].

o Actionable Insights: Effective feature engineering methods enable the extraction of actionable insights from raw data, facilitating more accurate and reliable predictions.

3. Dimensionality Reduction:

o Managing Complexity: Techniques like PCA and t-SNE help in reducing the complexity of high-dimensional datasets while retaining essential information. This simplification is crucial for maintaining model efficiency and preventing overfitting.

o Efficiency and Scalability: Dimensionality reduction techniques contribute to the development of more efficient and scalable predictive models, as seen in studies by [29] and [28].

Overall, the integration of meticulous data cleaning, strategic feature engineering, and advanced dimensionality reduction techniques consistently correlates with higher model accuracy across the studies. Employing these methods effectively ensures that the data used for model training is robust, relevant, and manageable, leading to more reliable and precise outcomes. In conclusion, optimizing data preprocessing is crucial for accurate and efficient water potability prediction, especially in settings with limited computational resources. This research contributes to making predictive modelling more accessible and applicable in diverse contexts, ensuring reliable and precise outcomes for water potability assessment.

5.2. Review in relation to the study objectives

Table 6 presents the contribution of research papers reviewed that mostly contribute to the research questions of the research study as presented in section 1.2 and Figure 6 presents the histogram of most papers contributing to the research questions.

Table 6:Research papers reviewed that mostly contribute to the research questions.

Objective	Study	Key Contributions
How does data cleaning impact the reliability and accuracy of machine learning models for water potability prediction?	[1]	Importance of understanding contaminants for data preprocessing.
	[8]	Clean data for precise assessments protecting vulnerable populations.
	[11]	Role of software engineering in maintaining clean datasets.
	[22]	Various data cleaning techniques and their impact on prediction models.
	[25]	Addressing the gap in water quality analysis with clean data.
	[30]	Vigilance in data cleaning for reliable human consumption analysis.
How does feature engineering contribute to improving the discriminative power of water potability datasets and enhance model interpretability?	[3]	Feature engineering's role in improving assessment accuracy.
	[6]	AI to process electronic nose data, enhancing feature extraction for contaminant detection.
	[5]	Optimized sensor placement using engineered features.

	[10]	Capillary electrophoresis evaluation with engineered features for better monitoring.
	[12]	Non-target analysis application, highlighting feature engineering's role in complex data handling.
	[23]	Machine learning techniques that rely on effective feature engineering for water infrastructure integrity.
	[28]	Insights into wastewater treatment processes improved by feature engineering.
	[29]	Application of clustering for condition-based maintenance highlighting feature engineering.
	[29]	Review of clustering applications for maintenance processes emphasizing feature extraction.
How do dimensionality reduction methods manage high-dimensional data while preserving essential information for accurate model predictions?	[2]	Deep learning benefiting from dimensionality reduction for precise water quality management.
	[9]	Enhanced real-time monitoring capabilities through dimensionality reduction.
	[13]	Continuous monitoring systems leveraging reduced dimensions for better accuracy.
	[14]	Nanosensors benefiting from reduced data complexity for real-time analysis.
	[15]	IoT integration with reduced data dimensions for smart water quality monitoring.
	[16]	Real-time monitoring improvements through effective dimensionality reduction.
	[21]	Portable biological spectroscopy enhanced by dimensionality reduction techniques.
	[26]	IoT innovations utilizing dimensionality reduction for efficient water quality monitoring.

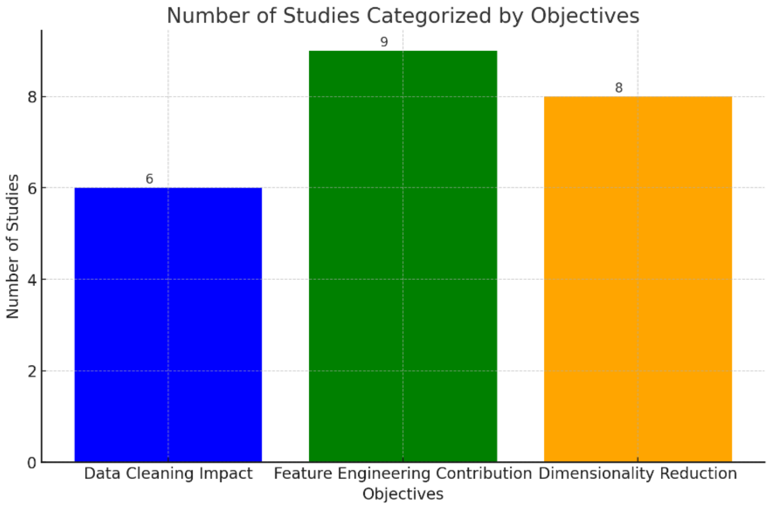

Figure 6:Histogram of papers contributing mostly to the research questions.

5.3. Implications for practice and policy.

The reviewed papers on Table 7 suggest several practical and policy implications to enhance water quality monitoring and treatment. For practice, there is an emphasis on adopting advanced technologies, integrating data-driven methodologies, and ensuring comprehensive real-time monitoring. Policymakers are encouraged to establish regulations and standards that promote these advanced practices, provide funding for research and development, and support the integration of health and environmental data. These measures aim to improve water quality management, ensure sustainable reclamation practices, and enhance the safety and reliability of potable water supplies.

Table 7:Several practical and policy implications to enhance water quality monitoring and treatment.

Paper	Implications for Practice	Implications for Policy
[9]	• Implement systematic approaches for reclamation of mining-degraded areas. • Use advanced technologies and best practices for effective reclamation.	• Develop and enforce regulations for reclamation of mining sites. • Provide incentives for adopting sustainable reclamation practices.
[28]	• Employ data-driven methodologies for treating sulfur-containing wastewater. • Integrate machine learning models to optimize treatment processes.	• Establish guidelines for using data analytics in wastewater treatment. • Support research in advanced data-driven treatment technologies.
[31]	• Standardize clustering methodologies for condition-based maintenance. • Develop adaptive algorithms for managing time-varying processes.	• Create policies promoting the use of advanced clustering techniques in maintenance. • Fund research for developing adaptive maintenance technologies.
[30]	• Integrate surveillance data with health outcomes to improve water quality monitoring. • Implement real-time monitoring technologies for timely interventions.	• Formulate policies that require the integration of health data with water quality monitoring. • Provide funding for the development of real-time monitoring systems.
[4]	• Use electronic nose systems for real-time monitoring of potable water quality. • Conduct extensive field trials to validate technology.	• Establish standards for the deployment of electronic nose systems. • Support policies that fund field trials and real-world applications.
[1]	• Focus on removing emerging contaminants in water purification processes. • Conduct socio-economic analyses of purification technologies.	• Develop regulations addressing emerging contaminants. • Provide economic incentives for advanced water purification technologies.
[5]	• Develop dynamic methodologies for placing water monitoring stations. • Utilize real-time data and predictive analytics for contamination detection.	• Formulate policies for adaptive placement of monitoring stations. • Encourage the use of predictive analytics in water quality management.
[6]	• Use interpretable models for monitoring cyanobacteria in potable water. • Validate models extensively in real-world settings.	• Establish guidelines for the use of black-box and interpretable models in water monitoring. • Support policies that fund real-world validation studies.
[7]	• Integrate multiple intelligent techniques for comprehensive water quality monitoring.	• Develop policies encouraging the integration of various intelligent monitoring techniques.

	• Transition experimental techniques to practical applications.	• Provide funding for the practical implementation of intelligent monitoring systems.

5.4. Final thoughts on the potential of AI in enhancing water potability.

AI has the potential to transform the water industry, enhancing the quality and safety of potable water through advanced monitoring, data analysis, and optimization techniques. This by addressing the challenges and fostering collaboration between technologists, policymakers, and water management authorities, AI can play a pivotal role in ensuring safe and reliable drinking water for all. The potential of AI in enhancing water potability, as evidenced by the papers provided, is substantial and multifaceted. Here are the key insights and detailed reasons from the papers presented in Table 8:

Table 8:Key insights and detailed reasons from the reviewed papers

Paper	Detailed Reasons
[1]	This paper discusses the nature and purification of potable water, emphasizing traditional methods. While it does not focus on AI directly, the foundational understanding of water contaminants and purification processes provides a baseline for integrating advanced AI-driven methods for more efficient and accurate water purification.
[2]	This review highlights the state-of-the-art applications of deep learning in water quality management. AI techniques, particularly deep learning, offer significant advancements in monitoring and predicting water quality parameters. They enhance the accuracy of detecting contaminants and predicting future water quality issues, enabling proactive management and intervention.
[3]	The systematic literature analysis on AI for surface water quality monitoring and assessment demonstrates that AI algorithms, including machine learning and neural networks, improve the precision of water quality assessment. These methods can analyze complex data sets from various sensors to provide real-time monitoring and early detection of pollution events.
[4]	This study introduces an electronic nose system for monitoring potable water quality. AI is used to process the data from the electronic nose, which mimics the human olfactory system. The AI-driven system can detect and identify different water contaminants, providing a rapid and reliable method for continuous water quality monitoring.
[5]	This paper focuses on methodologies for locating monitoring stations to detect contamination in potable water distribution systems. AI can enhance these methodologies by optimizing the placement of sensors and predicting potential contamination points based on historical and real-time data.
[6]	The study applies black-box modeling to electronic nose data for monitoring cyanobacteria in potable water. AI, through system identification techniques, helps in understanding and predicting the behavior of water quality parameters influenced by cyanobacteria, leading to better management strategies
[7]	A survey on intelligent techniques for potable water quality monitoring shows that AI techniques, such as machine learning and IoT (Internet of Things), significantly improve the detection, prediction, and management of water quality. AI systems can handle large volumes of data from diverse sources, providing comprehensive insights into water quality.
[8]	The study on potable water quality monitoring in primary schools in Bangladesh highlights the health risks associated with poor water quality. AI-driven analysis can enhance the monitoring and mitigation strategies by providing accurate and timely assessments, thus protecting vulnerable populations
[9]	This paper presents a near real-time monitoring system for N-nitrosodimethylamine in potable water using advanced chromatography techniques. AI algorithms can enhance the data analysis process, improving the speed and accuracy of contaminant detection.
[10]	The evaluation of capillary electrophoresis for monitoring ionic contaminants in space missions shows the potential for AI to improve the analysis and management of water quality in extreme environments, ensuring safe drinking water for astronauts.
[11]	AI applications in software engineering are discussed, highlighting the potential for integrating AI techniques in water quality monitoring systems to enhance their efficiency and accuracy.

[12]	The application of non-target analysis with LC-HRMS for monitoring water quality illustrates how AI can process complex analytical data to identify and quantify a wide range of contaminants, ensuring comprehensive water quality assessment.
[13]	The study on continuous active monitoring to identify cross-connections between potable water and effluent systems emphasizes the role of AI in providing continuous, real-time monitoring and alerting systems to prevent contamination.
[14]	The use of nanosensors for water quality monitoring demonstrates how AI can enhance the sensitivity and specificity of these sensors, providing precise and real-time water quality data.
[15]	The paper on IoT-based smart water quality monitoring outlines the integration of AI with IoT devices to provide real-time, accurate water quality data, improving domestic water quality management.
[16]	The use of chemical sensors for real-time water quality monitoring showcases the role of AI in processing sensor data, leading to timely and accurate water quality assessments.

Overall, these papers collectively highlight that AI, through various techniques such as machine learning, deep learning, and IoT integration, significantly enhances the monitoring, assessment, and management of potable water quality. AI provides real-time, accurate, and comprehensive insights, enabling proactive measures to ensure safe and clean drinking water.

References

1. Hussam, Abul (2013). Potable Water: Nature and Purification. Monitoring Water Quality: Pollution Assessment, Analysis, and Remediation, 261-283, ISSN 9780444593955, Elsevier Inc. https://doi.org/10.1016/B978-0-444-59395-5.00011-X
2. Wai, KP, Chia, MY, Koo, CH, Huang, YF, & Chong, WC (2022). Applications of deep learning in water quality management: A state-of-the-art review. Journal of Hydrology, Elsevier, 613, 128332. https://www.sciencedirect.com/science/article/pii/S0022169422009040
3. Ighalo, JO, Adeniyi, AG, & Marques, G (2021). Artificial intelligence for surface water quality monitoring and assessment: A systematic literature analysis. Modeling Earth Systems and Environment, Springer, 7(2),669-681. https://doi.org/10.1007/s40808-020-01041-z
4. Gardner, Julian W, Shin, Hyun Woo, Hines, Evor L, & Dow, Crawford S (2000). An electronic nose system for monitoring the quality of potable water. Sensors and Actuators B.69,336-341.www.elsevier.nlrlocatersensorb
5. Chastain, James R, & Asce, M. Methodology for Locating Monitoring Stations to Detect Contamination in Potable Water Distribution Systems. Journal of infrastructure systems,12(4),252-259. https://doi.org/10.1061/ASCE1076-0342200612:4252
6. Searle, G.E., Gardner, J.W., Chappell, M.J., Godfrey, K.R., & Chapman, M.J. (2000). System Identification of Electronic Nose Data for Monitoring Cyanobacteria in Potable Water: Black-Box Modeling. IFAC Proceedings Volumes, 33(15), 145-150. Elsevier BV, https://doi.org/10.1016/s1474-6670(17)39741-0
7. Ntshako, Neo, Markus, Elisha Didam, & Abu-Mahfouz, Adnan M. (2019). Potable Water Quality Monitoring: A Survey of Intelligent Techniques. International Multidisciplinary Information technology (IMITEC) (pp.1-6). IEE.
8. Rahman, Aminur, Hashem, Abul, & Nur-A-Tomal, Shahruk (2016). Potable water quality monitoring of primary schools in Magura district, Bangladesh: childrenâ€™s health risk assessment. Environmental Monitoring and Assessment, 188(12), Springer International Publishing, https://doi.org/10.1007/s10661-016-5692-6
9. Fujioka, Takahiro, Tanisue, Taketo, Roback, Shannon L., Plumlee, Megan H., Ishida, Kenneth P., & Kodamatani, Hitoshi (2017). Near real-time N -nitrosodimethylamine monitoring in potable water reuse via online high-performance liquid chromatography-photochemical reaction-chemiluminescence. Environmental Science: Water Research and Technology, 3(6), 1032-1036, Royal Society of Chemistry. https://doi.org/10.1039/c7ew00296c

10. Mudgett, Paul D, Schultz, John R, & Sauer, Richard L.(1992). Evaluation of Capillary Electrophoresis for In-flight Ionic Contaminant Monitoring of SSF Potable Water.SAE Transactions, 888-897.
11. Mata, Mirna A. MunÌƒoz, & Engineers, Institute of Electrical and Electronics Applications in Software Engineering. (2019. proceedings of the 8th International Conference on Software Process Improvement: Guadalajara, Jalisco, MeÌxico, October 23-25, 2019., ISSN 9781728155555.
12. Bader, Tobias, Schulz, Wolfgang, & Lucke, Thomas (2016). Application of non-target analysis with LC-HRMS for the monitoring of raw and potable water: Strategy and results. ACS Symposium Series, 1242, 49-70, ISSN 9780841231955, American Chemical Society. https://doi.org/10.1021/bk-2016-1242.ch003
13. Friedler, E., Alfiya, Y., Shaviv, A., Gilboa, Y., Harussi, Y., & Raize, O. (2015). A continuous active monitoring approach to identify cross-connections between potable water and effluent distribution systems. Environmental Monitoring and Assessment, 187(3), Kluwer Academic Publishers. https://doi.org/10.1007/s10661-015-4350-8
14. Vikesland, PJ (2018). Nanosensors for water quality monitoring. Nature nanotechnology,13(8), 651-660. https://www.nature.com/articles/s41565-018-0209-9
15. Jan, F, Min-Allah, N, & Düştegör, D (2021). Iot based smart water quality monitoring: Recent techniques, trends and challenges for domestic applications. https://www.mdpi.com/2073-4441/13/13/1729
16. Yaroshenko, I, Kirsanov, D, Marjanovic, M, Lieberzeit, PA, & ... (2020). Real-time water quality monitoring with chemical sensors. Sensors, 20(12), 20.12:3432. https://www.mdpi.com/1424-8220/20/12/3432
17. Syeed, MMM, Hossain, MS, Karim, MR, Uddin, MF, & ... (2023). Surface water quality profiling using the water quality index, pollution index and statistical methods: A critical review. Environmental and Sustainability Indicators, 18,100247. https://www.sciencedirect.com/science/article/pii/S2665972723000247
18. Ez-Zaouy, Y, Bouchaou, L, Saad, A, Hssaisoune, M, & ... (2022). Morocco's coastal aquifers: Recent observations, evolution and perspectives towards sustainability. Environmental Pollution, 293,118498. https://www.sciencedirect.com/science/article/pii/S0269749121020807
19. Tashtoush, B, & Alshoubaki, A (2023). Atmospheric water harvesting: A review of techniques, performance, renewable energy solutions, and feasibility. Energy, Elsevier, https://www.sciencedirect.com/science/article/pii/S0360544223015803
20. Khorramifar, A, Karami, H, Lvova, L, Kolouri, A, Łazuka, E, & ... (2023). Environmental engineering applications of electronic nose systems based on MOX gas sensors. sensors, 23(12),5716. https://www.mdpi.com/1424-8220/23/12/5716
21. Damit, B, & Antoine, M (2021). Portable biological spectroscopy: Field applications. Portable spectroscopy and spectrometry,545-563. https://doi.org/10.1002/9781119636489.ch22
22. Mittal, A, Patwal, S, Adhikari, M, & ... (2023). A Review of Various Water Quality Prediction Models and Techniques. 2023 5th International Conference on Inventive Research in Computing Applications. (pp.614-620). IEEE. https://ieeexplore.ieee.org/abstract/document/10220687/
23. García, J, Leiva-Araos, A, Diaz-Saavedra, E, Moraga, P, & ... (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://www.mdpi.com/2076-3417/13/22/12497
24. Ooko, SO, Pamela, EK, & Kwagalakwe, G (2023). Use of Machine Learning for Realtime Water Quality Prediction. 2023 IEEE AFRICON,1-6. https://ieeexplore.ieee.org/abstract/document/10293701/
25. Flores, RR, & Domínguez, MJ (2023). Contributions to Reduce the Gap on Water Quality Analysis in Chile and Latin America: State of the Art., In 2023 IEEE CHILEAN Conference on

Electrical, Electronics Engineering, Information and Communication Technologies (CHILEON)(pp.1-6).https://easychair.org/publications/preprint_download/nr4w

26. Alshami, A, Ali, E, Elsayed, M, Eltoukhy, AEE, & ... (2024). IoT Innovations in Sustainable Water and Wastewater Management and Water Quality Monitoring: A Comprehensive Review of Advancements, Implications, and Future Directions. IEEE Access, https://ieeexplore.ieee.org/abstract/document/10506924/
27. Júnior, AP, Freitas, BG, Oliveira, R de Souza, & ... (2022). Systematic review on the reclamation of areas degraded by mining. Research, Society and Development,11(8),e2711830706-e2711830706. https://rsdjournal.org/index.php/rsd/article/view/30706
28. Jin, L, Lu, J, Sun, X, Huang, H, & Ren, H (2023). Data-driven insights into treatment of sulfur-containing organic wastewater. Journal of Cleaner Production, 139878. https://www.sciencedirect.com/science/article/pii/S0959652623040362
29. Quatrini, E, Colabianchi, S, Costantino, F, & Tronci, M (2022). Clustering application for condition-based maintenance in time-varying processes: A review using latent dirichlet allocation. Applied Sciences,12(2) 814. https://www.mdpi.com/2076-3417/12/2/814
30. Lopes, RH (2022). Vigilância da qualidade da água para consumo humano: scoping review. https://repositorio.ufrn.br/handle/123456789/51974
31. Quatrini, E, Colabianchi, S, Costantino, F, & Tronci, M (2021). Clustering application for condition-based maintenance in time-varying processes: a review. preprints.org, https://www.preprints.org/manuscript/202111.0089