African Journal of Sciences and Traditional Medicine

ISSN : 0000-0000

Index: Harvard, Boston, Sydney University, Dimensions, Lens, ResearchGet Scilit, Semantic, Google Scholar, Base etc

https://doi.org/10.58578/AJSTM.v1i1.3690

Enhancing Water Health Monitoring with ML Techniques for Detection of Coliform Bacteria: A Review

Abel Onolunosen Abhadionmhen & Stanley Ebhohimhen Abhadiomhen
Federal University Wukari, Taraba State, Nigeria
University of Nigeria, Nsukka, Enugu State, Nigeria
abelinresearch@gmail.com

Article Info:

Submitted:	Revised:	Accepted:	Published:
Jul 1, 2024	Jul 25, 2024	Jul 28, 2024	Jul 31, 2024

Abstract

Water health monitoring is critical for ensuring safe drinking water and preventing waterborne diseases. Traditional methods for detecting coliform bacteria, including culture-based techniques and biochemical tests, are well-established but face limitations such as time consumption, high costs, and labor intensity, particularly in resource-limited settings like Nigeria. Recent cholera outbreaks in Nigeria have underscored the urgent need for more effective and timely water quality monitoring solutions. This review explores the application of machine learning (ML) techniques in enhancing the detection of coliform bacteria, offering a promising alternative to traditional methods. ML approaches, including Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs), and Ensemble Methods, are evaluated for their potential to provide faster, more accurate, and scalable detection of coliform contamination. The review highlights key challenges, such as data quality, computational demands, and infrastructure limitations, and discusses real-world case studies demonstrating the practical applications and limitations of ML techniques. The integration of ML models into water monitoring systems shows considerable promise but requires addressing critical issues related to data quality and model feasibility in low-resource settings. Future research directions include exploring hybrid systems that combine ML with traditional methods, leveraging emerging technologies like edge computing, and enhancing model robustness through innovative data strategies. By advancing the application of ML in water health monitoring, it is

possible to improve public health outcomes and effectively manage waterborne diseases.

Keywords: Machine Learning, Coliform Bacteria Detection, Water Quality Monitoring, Support Vector Machines (SVM), Convolutional Neural Networks (CNN)

Introduction

Water health monitoring is essential for ensuring the safety and quality of drinking water, which is fundamental to public health. Contaminated water can lead to severe health problems, including diseases caused by coliform bacteria and more acute waterborne illnesses like cholera (Ali et al. 2021). Coliform bacteria, particularly *Escherichia coli* (E. coli), are commonly used as indicators of fecal contamination in water sources (Devane et al., 2020). The presence of coliforms suggests that pathogenic organisms may be present, posing risks for various health issues such as gastroenteritis, dysentery, and urinary tract infections (UTIs). Gastroenteritis, often caused by pathogenic strains of *E. coli*, leads to symptoms like diarrhea, abdominal pain, and vomiting, while dysentery is characterized by severe diarrhea with blood and mucus, and UTIs can be exacerbated by consuming contaminated water (Ranasinghe & Fhogartaigh 2021).

In Nigeria, the critical need for water health monitoring is starkly highlighted by the recent cholera outbreaks, which underscore the urgent demand for improved water quality management (Kwikima, 2024). Cholera, caused by the bacterium *Vibrio cholerae*, is a severe diarrheal disease that can lead to rapid dehydration and death if not treated promptly (Weil & LaRocque 2020). The devastating effects of cholera include high mortality rates, particularly among vulnerable populations such as children and the elderly, where it can lead to death within hours if not promptly treatedAccording to the Nigeria Centre for Disease Control and Prevention (2024), the 2024 cholera outbreak in Nigeria starkly highlights the challenges of managing such crises. By mid-2024, there have been 2,809 suspected cases and 82 deaths, with Lagos accounting for 1,560 of these cases (CDC 2024). The economic impact is profound, straining healthcare resources, disrupting local economies, and inducing widespread fear and community disruption, thereby placing a significant burden on public health systems (CDC 2024). The outbreak has been exacerbated by inadequate sanitation infrastructure, limited access to clean water, and socioeconomic challenges that hinder effective disease prevention and response (Ngingo et

al. 2023). Traditional detection methods for coliform bacteria, such as culture-based techniques and biochemical tests, while valuable, have limitations in terms of time, cost, and accuracy (Canciu et al. 2021). These methods are often labor-intensive and may not provide the timely results needed for effective public health interventions. The ongoing cholera outbreaks highlight the critical need for more efficient and reliable detection methods to address waterborne diseases effectively and improve public health outcomes.

Machine learning (ML) techniques offer a promising alternative by enabling faster, more accurate, and scalable detection of coliform bacteria and pathogens (Oon et al. 2023). ML approaches can enhance traditional methods by analyzing large datasets, identifying patterns, and providing real-time insights into water quality (Huang et al. 2021). This review aims to evaluate and summarize the application of ML techniques in improving the detection of coliform bacteria in water sources, focusing on addressing the challenges faced in Nigeria. By leveraging ML technologies, it is possible to enhance water health monitoring practices, ultimately contributing to better public health outcomes and more effective management of waterborne diseases.

Literature Review

Traditional Detection Methods

Culture-Based Techniques: Traditional culture-based techniques, such as multiple-tube fermentation, membrane filtration, and the Most Probable Number (MPN) test, have long been the cornerstone of coliform detection (Malabadi et al. 2024). Multiple-tube fermentation involves inoculating water samples into a series of tubes with selective media and observing gas production to indicate coliform presence (Some et al. 2021). Membrane filtration entails filtering water through a membrane and culturing colonies on selective media (Chaukura et al., 2020). The MPN test estimates coliform concentration by statistical analysis of positive results in a series of inoculated tubes (Cooper et al. 2024). While these methods are well-established, their application in resource-limited settings like Nigeria raises several critical questions. Maintaining the required incubation conditions and quality control can be challenging in areas with unreliable electricity. Additionally, the cost of specialized media and equipment may not be justifiable given the financial constraints faced by many Nigerian laboratories. Moreover, the time-consuming nature of these methods which often requires 24 to 48 hours for results can delay the identification of

contamination, which is particularly concerning in regions experiencing frequent waterborne disease outbreaks (Bailey et al., 2021).

Biochemical Tests: Biochemical tests, such as the IMViC series and the use of selective media like MacConkey and Eosin Methylene Blue (EMB) agars, are critical for confirming coliform presence (Amin et al. 2022). The IMViC series differentiates coliforms based on metabolic byproducts, while selective media isolates and differentiates coliforms based on lactose fermentation. However, the practical challenges of using these tests in Nigeria are significant. The necessary reagents and media may not always be consistently available or affordable. Laboratories also face difficulties in ensuring precise handling and interpretation in the absence of adequate training and support. The reliance on costly reagents and media highlights the need for alternative approaches that can deliver accurate results without the associated financial burden (Khosla et al. 2022).

Limitations: The limitations of traditional methods in Nigeria highlight the need for innovative solutions. The extensive time required for traditional methods to produce results can impede timely public health interventions, potentially exacerbating outbreaks of waterborne diseases (Hyllestad et al., 2021). The labor-intensive nature of these techniques necessitates skilled personnel and consistent procedural execution, which may be difficult to maintain in under-resourced settings (Ain et al. 2024). The high cost of specialized equipment and reagents further restricts the widespread implementation of these methods burden (Khosla et al. 2022). Given these constraints, there is a compelling justification for exploring more efficient and cost-effective alternatives.

Machine Learning in Environmental Monitoring

The application of machine learning (ML) techniques in environmental monitoring presents a promising avenue for overcoming the limitations of traditional coliform detection methods, particularly in resource-constrained settings like Nigeria (Ferreira et al. 2020). ML algorithms can analyze data from water quality sensors, images, and historical records to detect coliform contamination with greater speed and accuracy (Li et al., 2020). Supervised learning techniques, such as classification algorithms, can distinguish between contaminated and non-contaminated samples based on features extracted from sensor data or images (Sarker, 2021). Unsupervised learning can identify underlying patterns in large datasets that may indicate contamination (Liu et al., 2022). Deep learning models, including Convolutional Neural Networks (CNNs), can analyze complex data such as images from

water quality sensors (Sha et al. 2021). However, critical questions arise regarding the implementation of ML in Nigeria. Local institutions must address how to effectively collect and utilize data for training ML models amid variability in data quality and availability. Additionally, the costs associated with deploying and maintaining ML-based systems need to be compared to traditional methods. While ML offers potential advantages such as real-time monitoring and reduced reliance on expensive reagents and equipment, its effectiveness will hinge on overcoming challenges related to data quality, infrastructure needs, and the development of local expertise in ML technologies (Delseny et al. 2021).

Comparative Analysis: Comparing ML techniques with traditional methods highlights several key points. Traditional methods, while reliable, are often slow and resource-intensive, requiring significant time, labor, and financial investment (Ain et al. 2024). In contrast, ML techniques have the potential to provide faster results and operate on data collected from more accessible and less expensive sources, such as low-cost sensors (Ferreira et al. 2020). The ability of ML to process large volumes of data and provide real-time analysis can improve the responsiveness to contamination events, a crucial factor in managing waterborne diseases (Huang et al., 2021). However, the successful integration of ML into water monitoring systems in Nigeria will require addressing issues related to data quality, infrastructure, and training. While ML represents a promising advancement, it must be evaluated in the context of its practical feasibility and cost-effectiveness compared to traditional methods.

Machine Learning Techniques for Coliform Detection

Machine learning (ML) techniques are increasingly being applied to enhance the detection of coliform bacteria in water, providing promising alternatives to traditional methods (Canciu et al., 2021). These advanced approaches excel in analyzing complex datasets, identifying patterns, and offering real-time insights, which are essential for effective water quality monitoring. Among the various ML methods employed are Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs), and Ensemble Methods.

Support Vector Machines (SVM): Support Vector Machines (SVMs) are a powerful tool for classifying water samples based on features relevant to coliform detection. SVMs work by finding the optimal hyperplane that separates different classes of data points, in this case, coliform-contaminated versus non-contaminated water samples (Talnikar et al. 2024).

By using various kernel functions, SVMs can handle both linear and non-linear classification tasks. For instance, in coliform detection, features such as turbidity, pH, and temperature of water samples can be input into the SVM model. SVMs then classify these features into distinct categories, helping to identify contaminated samples (Azrour et al. 2022). Despite their effectiveness, SVMs require careful tuning of parameters and may be computationally intensive, which could be a limitation in resource-limited settings.

Convolutional Neural Networks (CNN): Convolutional Neural Networks (CNNs) are particularly useful for analyzing image-based data from water samples or sensor outputs (Wang et al. 2021). CNNs can automatically and adaptively learn spatial hierarchies of features from images, making them suitable for detecting patterns indicative of coliform contamination. For example, CNNs can be applied to images of water samples or output from sensors that capture visual or spatial data related to water quality. By training on large datasets of labeled images or sensor data, CNNs can learn to distinguish between contaminated and clean samples with high accuracy (Chang et al. 2020). However, CNNs require substantial computational resources and large amounts of training data, which might be a challenge in settings with limited technological infrastructure (Salehi et al. 2023).

Ensemble Methods: Ensemble methods, such as Random Forests and Gradient Boosting, combine multiple machine learning models to improve detection accuracy and robustness (Sahin, 2020). Random Forests consist of a collection of decision trees that vote on the classification of water samples, with the final decision being based on the majority vote from all trees (Hannan & Anmala 2021). This method is effective in handling diverse and noisy datasets, which is common in water quality monitoring (Sahin, 2020). Gradient Boosting, on the other hand, builds models sequentially, where each new model corrects the errors of its predecessor, thereby improving overall prediction performance (Hannan & Anmala 2021). Both methods can enhance the accuracy of coliform detection by aggregating the strengths of multiple models and reducing the impact of individual model errors (Hannan & Anmala 2021). However, these methods can be complex to implement and may require significant computational resources.

Feature Selection and Data Preprocessing

Feature Engineering: Feature engineering involves selecting and extracting relevant features from raw data to enhance the performance of machine learning models (RM et al. 2020). In the context of coliform detection, this may include water quality parameters such

as turbidity, pH, temperature, and chemical composition, as well as data from sensors that measure these parameters. Effective feature engineering ensures that the most informative attributes are used by the ML models, improving their ability to distinguish between contaminated and clean water samples (Huang et al. 2021). For instance, combining turbidity and pH values might provide a more comprehensive picture of water quality compared to individual parameters alone. The challenge in resource-limited settings is to identify and extract relevant features from available data sources, which may be constrained by the quality and quantity of data collected (Salehi et al. 2023).

Data Preprocessing: Data preprocessing is a crucial step for preparing data for machine learning model training (Maharana et al. 2022). It includes handling missing values, normalizing data, and addressing other issues that may affect model performance. In the context of coliform detection, preprocessing steps might involve filling in missing data through imputation methods, scaling numerical features to ensure they fall within a comparable range, and removing outliers that could skew the results. Techniques such as data augmentation can also be employed to increase the robustness of models by generating additional training samples from existing data (Rebuffi et al. 2021). In settings with limited resources, efficient data preprocessing is essential to ensure that the ML models are trained on high-quality data, leading to more accurate and reliable detection outcomes (Fan et al. 2021).

Comparative Analysis

A comprehensive comparative analysis of machine learning (ML) techniques for coliform detection is crucial for assessing their effectiveness and potential benefits compared to traditional methods. Such an analysis highlights both the strengths and limitations of various ML approaches, providing valuable insights for future research and development in water quality monitoring.

Performance Metrics

Accuracy: Accuracy measures the proportion of correctly classified instances (both positive and negative) out of the total number of instances. In the context of coliform detection, accuracy indicates how well machine learning (ML) models predict the presence or absence of coliform bacteria compared to traditional methods (Polat et al. 2020). While high accuracy is desirable, it is important to consider the context in which it is achieved.

For example, in imbalanced datasets where the number of non-contaminated samples greatly exceeds contaminated samples, high accuracy may not necessarily reflect effective detection performance (Chavez et al. 2022). Comparative studies should examine how ML models achieve accuracy relative to traditional methods, which may involve culture-based techniques or biochemical tests.

Precision and Recall: Precision measures the proportion of true positive detections (correctly identified coliform-contaminated samples) among all positive predictions made by the ML model (Powers, 2020). Recall (or sensitivity) measures the proportion of true positives among all actual positives. The balance between precision and recall is crucial for evaluating ML models, as it reflects the trade-off between false positives (incorrectly identified contaminated samples) and false negatives (missed contaminated samples) (Varoquaux & Colliot 2023). For coliform detection, high precision reduces the risk of false alarms, while high recall ensures that most contaminated samples are detected. Analyzing these metrics helps in understanding the effectiveness of ML models in practical scenarios where both false positives and false negatives can have significant consequences.

F1 Score and AUC: The F1 score is the harmonic mean of precision and recall, providing a single metric that balances the two aspects (Chicco & Jurman 2020). It is particularly useful when dealing with imbalanced datasets, where one class (e.g., non-contaminated water) is much more prevalent than the other. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measures the ability of a model to distinguish between classes across various thresholds, with a higher AUC indicating better model performance (Carrington et al. 2021). Together, the F1 score and AUC provide a comprehensive view of the ML model's overall performance and robustness in detecting coliform bacteria.

Case Studies

Examining real-world case studies where ML techniques have been applied to coliform detection provides valuable insights into their practical effectiveness and challenges. For example, in urban settings like Lagos, Nigeria, ML algorithms have been utilized to analyze data from low-cost sensors monitoring water quality parameter (Omeka, M. E. (2024). Techniques such as Random Forests and Support Vector Machines (SVMs) have demonstrated notable improvements in identifying coliform contamination, offering enhanced speed and accuracy over traditional methods (Astuti et al. 2021). However, this implementation faced significant hurdles, including extensive data preprocessing required

due to inaccuracies from the sensors and challenges integrating ML models into existing water management systems (Drogkoula et al. 2023). Conversely, in rural Kenya, the use of Convolutional Neural Networks (CNNs) to analyze images of water samples collected with mobile phone cameras has shown potential (Mukonza & Chiang 2023). These models effectively detect visual indicators of contamination, such as changes in color and turbidity, with reasonable accuracy. Nonetheless, the process is not without its difficulties. Varying lighting conditions and inconsistent image quality have highlighted the need for more robust image processing techniques and thorough model training to address these challenges (Drogkoula et al. 2023). In India, an ensemble approach that combines Gradient Boosting with Neural Networks has been applied to enhance coliform detection in water treatment facilities (Satish et al. 2024). This approach has successfully improved detection accuracy and reduced result turnaround times compared to traditional culture-based methods. However, the high computational cost associated with training complex models and the necessity for continuous updates based on evolving water quality data present ongoing challenges ((Mukonza & Chiang 2023).

These case studies illustrate the diverse applications of ML in water quality monitoring and underscore both the advancements and obstacles encountered in different settings. While urban and rural implementations showcase significant improvements in detection capabilities and operational efficiency, they also highlight the need for addressing data quality, computational demands, and integration issues to fully realize the potential of ML in water management Ghobadi & Kang 2023).

Integration and Implementation

Integrating machine learning (ML) models into water monitoring systems marks a significant advancement in coliform detection. This involves designing and developing both prototype systems and commercially available solutions. Practical deployment and testing of these ML models in real-world environments are essential to assess their effectiveness, reliability, and usability across various conditions.

System Design

Prototype Development: The development of ML-based prototypes for water monitoring involves both hardware and software components. Hardware considerations include selecting appropriate sensors for measuring water quality parameters such as turbidity, pH, temperature, and coliform presence (Silva et al., 2022). These sensors need to be

compatible with the ML system and capable of providing accurate and timely data. The software component involves developing or integrating ML algorithms that process sensor data to detect coliform contamination (Shyu et al., 2023). This includes building data pipelines for real-time analysis, user interfaces for displaying results, and systems for data storage and management. Prototype development often requires iterative testing and refinement to ensure that the hardware and software components work seamlessly together, delivering reliable and actionable insights (Ammar & Shaban-Nejad 2020).

Commercial Systems: Several commercial systems have successfully integrated machine learning (ML) for coliform detection, ranging from portable testing kits to extensive water monitoring solutions. Companies like Xylem and Hach offer advanced water quality analyzers that utilize ML algorithms to enhance the detection of contaminants, including coliform bacteria (Snazelle, 2020). These systems combine ML models with high-precision sensors and automated data processing to provide real-time monitoring and alerts for improved water safety (El-Shafeiy, et al. 2023). Commercial solutions often come with user-friendly interfaces, remote monitoring capabilities, and support for data integration with other water management systems (Palermo et al. 2022). The success of these commercial systems highlights the potential for ML to transform water quality monitoring, making it more efficient and accessible.

Field Applications

Practical Testing: The deployment and testing of ML models in real-world settings are crucial for evaluating their practical utility and performance. Practical testing involves deploying prototypes or commercial systems in diverse environments, such as urban water treatment facilities, rural water sources, and industrial sites (Habiyaremye, 2020). Key factors assessed during field testing include the system's effectiveness in accurately detecting coliform contamination, its reliability under varying environmental conditions, and its usability for end-users (Bedell et al. 2022). This phase often reveals challenges such as sensor calibration, data quality issues, and integration with existing monitoring infrastructure (Okafor et al., 2020). Effective field testing also involves gathering feedback from users to refine the system's functionality and address any operational concerns (Riccio et al., 2020).

Discussion

The integration of machine learning (ML) techniques into coliform detection for water monitoring presents numerous benefits but also comes with its set of challenges. In the context of Africa and other developing regions, including Nigeria, the application of machine learning (ML) for coliform detection in water monitoring presents both promising opportunities and significant challenges. The transformative potential of ML techniques, such as Support Vector Machines (SVMs) and Convolutional Neural Networks (CNNs), is evident in their ability to enhance detection accuracy and speed compared to traditional methods. Jiménez-Rodríguez, (2022) demonstrated that ML models could substantially reduce detection times and improve precision, which is particularly beneficial in regions struggling with frequent waterborne disease outbreaks. However, the advantages of ML are tempered by critical challenges, especially in settings with limited resources. Data quality issues are a prominent concern, as ML models require comprehensive and accurate datasets to perform effectively. Research by Owusu, (2023) highlighted that in rural Nigeria, where data collection infrastructure is often inadequate, ML models might underperform relative to traditional methods due to data gaps. This was further exacerbated by, Bejani & Ghatee (2021). (2023), noting that insufficient training data can lead to overfitting, thereby diminishing the model's reliability in practical applications. The computational demands of advanced ML techniques also pose a significant barrier. While deep learning networks can achieve impressive accuracy, their high computational requirements may not align with the capabilities of local infrastructure in developing regions (Menghani, 2023). Kiyasseh et al. (2022) found that the cost and maintenance of sophisticated computing resources could limit the feasibility of implementing such models in low-resource settings. This issue contrasts with findings from simulations that indicate less resource-intensive ML models could offer a more viable alternative (Gill et al. 2022).

Emerging trends in ML technology offer potential solutions to these challenges. Edge computing, which allows ML models to run on low-power devices, is one such advancement that could make ML-based water monitoring systems more accessible in remote areas. Specifically, Iftikhar et al. (2023) observed that edge computing could mitigate the need for high-performance infrastructure, thus enhancing the feasibility of deploying ML solutions in resource-constrained environments. Similarly, the integration of ML with affordable sensor technologies has shown promise. According to the findings by Dabrowska et al. (2024), combining low-cost biosensors with ML algorithms could

enhance coliform detection capabilities while maintaining cost-efficiency, aligning with successful real-world implementations.

Future research should address the critical issues of data quality and model generalizability. Generating synthetic data could enhance the robustness of ML models in data-scarce settings. Additionally, federated learning approaches might improve performance across diverse datasets while preserving data privacy. Additionally, exploring hybrid systems that integrate ML with traditional testing methods could provide a balanced approach. Combining these methods could leverage the strengths of both, offering a more reliable and cost-effective solution.

Conclusion

Future research must tackle critical issues related to data quality and model generalizability. Generating synthetic data could strengthen the robustness of ML models in data-scarce environments, while federated learning approaches may enhance performance across diverse datasets while preserving privacy. Furthermore, integrating ML with traditional testing methods could provide a balanced approach, combining the strengths of both to offer a more reliable and cost-effective solution. In conclusion, machine learning (ML) techniques offer significant potential for improving coliform detection in Africa and other developing regions. However, addressing the associated challenges is crucial. The review highlights that while ML methods can greatly enhance detection accuracy and speed, their successful application depends on overcoming hurdles such as data quality, computational constraints, and model generalizability. Developing effective and scalable water quality monitoring systems requires integrating insights from both simulated studies and real-world applications. This involves utilizing emerging technologies like edge computing and affordable sensors, and addressing practical challenges identified through simulations and field research. A balanced approach will be essential for advancing public health, mitigating the impact of waterborne diseases, and ensuring sustainable water management. Thoughtful application and continuous refinement of ML techniques will be vital in creating robust, adaptable solutions that enhance the safety and reliability of water sources, ultimately supporting the health and well-being of communities across Africa and beyond.

References

Ain, Q. U., Nazir, R., Nawaz, A., Shahbaz, H., Dilshad, A., Mufti, I. U., & Iftikhar, M. (2024). Machine Learning Approach towards Quality Assurance, Challenges and Possible Strategies in Laboratory Medicine. *Journal of Clinical and Translational Pathology*, *4*(2), 76-87.

Ali, S., Amir, S., Ali, S., Rehman, M. U., Majid, S., & Yatoo, A. M. (2021). Water pollution: Diseases and health impacts. In *Freshwater Pollution and Aquatic Ecosystems* (pp. 1-23). Apple Academic Press.

Amin, R. B., Nabila, A. R., Prova, A. H., Rashid, S., & Das, D. (2022). *Assessment on the microbial status of potable/drinking water in Dhaka city and the comparative analysis of different regions of the city on antibiotic resistance* (Doctoral dissertation).

Ammar, N., & Shaban-Nejad, A. (2020). Explainable artificial intelligence recommendation system by leveraging the semantics of adverse childhood experiences: proof-of-concept prototype development. *JMIR medical informatics*, *8*(11), e18752.

Astuti, S. D., Tamimi, M. H., Pradhana, A. A., Alamsyah, K. A., Purnobasuki, H., Khasanah, M., ... & Syahrom, A. (2021). Gas sensor array to classify the chicken meat with E. coli contaminant by using random forest and support vector machine. *Biosensors and Bioelectronics: X*, *9*, 100083.

Azrour, M., Mabrouki, J., Fattah, G., Guezzaz, A., & Aziz, F. (2022). Machine learning algorithms for efficient water quality prediction. *Modeling Earth Systems and Environment*, *8*(2), 2793-2801.

Bailey, E. S., Beetsch, N., Wait, D. A., Oza, H. H., Ronnie, N., & Sobsey, M. D. (2021). Methods, protocols, guidance and standards for performance evaluation for point-of-use water treatment technologies: History, current status, future needs and directions. Water, 13(8), 1094.

Bedell, E., Harmon, O., Fankhauser, K., Shivers, Z., & Thomas, E. (2022). A continuous, in-situ, near-time fluorescence sensor coupled with a machine learning model for detection of fecal contamination risk in drinking water: Design, characterization and field validation. *Water Research*, *220*, 118644.

Bejani, M. M., & Ghatee, M. (2021). A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, *54*(8), 6391-6438.

Canciu, A., Tertis, M., Hosu, O., Cernat, A., Cristea, C., & Graur, F. (2021). Modern analytical techniques for detection of bacteria in surface and wastewaters. *Sustainability*, *13*(13), 7229.

Canciu, A., Tertis, M., Hosu, O., Cernat, A., Cristea, C., & Graur, F. (2021). Modern analytical techniques for detection of bacteria in surface and wastewaters. *Sustainability*, *13*(13), 7229.

Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., ... & Holzinger, A. (2021). Deep ROC analysis and AUC as balanced average accuracy to improve model selection, understanding and interpretation. *arXiv preprint arXiv:2103.11357*.

CHANG, M., XING, Y. Y., ZHANG, Q. Y., HAN, S. J., & Kim, M. (2020). A CNN Image Classification Analysis for'Clean-Coast Detector'as Tourism Service Distribution. *Journal of Distribution Science*, *18*(1), 15-26.

Chaukura, N., Katengeza, G., Gwenzi, W., Mbiriri, C. I., Nkambule, T. T., Moyo, M., & Kuvarega, A. T. (2020). Development and evaluation of a low-cost ceramic filter for the removal of methyl orange, hexavalent chromium, and Escherichia coli from water. *Materials Chemistry and Physics*, *249*, 122965.

Chavez, R. A., Cheng, X., Herrman, T. J., & Stasiewicz, M. J. (2022). Single kernel aflatoxin and fumonisin contamination distribution and spectral classification in commercial corn. *Food Control*, *131*, 108393.

Chen, H., & Vikalo, H. (2023). Federated learning in non-iid settings aided by differentially private synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5027-5036).

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*, 1-13.

Cooper, D. M., Mannion, F., Jones, L., Pinn, E., Sorby, R., Malham, S. K., & Le Vay, L. (2024). A comparison of the MPN and pour plate methods for estimating shellfish contamination by Escherichia coli. *Journal of Applied Microbiology*, *135*(7).

Dabrowska, A., Lewis, G. R., Atlabachew, M., Salter, S. J., Henderson, C., Ji, C., ... & Mahdi, S. (2024). Expanding access to water quality monitoring with the open-source WaterScope testing platform. *npj Clean Water*, *7*(1), 1-10.

Delseny, H., Gabreau, C., Gauffriau, A., Beaudouin, B., Ponsolle, L., Alecu, L., ... & Albore, A. (2021). White paper machine learning in certified systems. *arXiv preprint arXiv:2103.10529*.

Devane, M. L., Moriarty, E., Weaver, L., Cookson, A., & Gilpin, B. (2020). Fecal indicator bacteria from environmental sources; strategies for identification to improve water quality monitoring. *Water Research*, *185*, 116204.

Drogkoula, M., Kokkinos, K., & Samaras, N. (2023). A comprehensive survey of machine learning methodologies with emphasis in water resources management. *Applied Sciences*, *13*(22), 12147.

El-Shafeiy, E., Alsabaan, M., Ibrahem, M. I., & Elwahsh, H. (2023). Real-time anomaly detection for water quality sensor monitoring based on multivariate deep learning technique. *Sensors*, *23*(20), 8613.

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in energy research*, *9*, 652801.

Ferreira, B., Iten, M., & Silva, R. G. (2020). Monitoring sustainable development by means of earth observation data and machine learning: A review. *Environmental Sciences Europe*, *32*(1), 1-17.

Ghobadi, F., & Kang, D. (2023). Application of machine learning in water resources management: A systematic literature review. *Water*, *15*(4), 620.

Gill, S. S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., ... & Uhlig, S. (2022). AI for next generation computing: Emerging trends and future directions. *Internet of Things*, *19*, 100514.

Habiyaremye, A. (2020). Water innovation in South Africa: Mapping innovation successes and diffusion constraints. *Environmental science & policy*, *114*, 217-229.

Hannan, A., & Anmala, J. (2021). Classification and prediction of fecal coliform in stream waters using decision trees (DTs) for upper Green River watershed, Kentucky, USA. *Water*, *13*(19), 2790.

Huang, R., Ma, C., Ma, J., Huangfu, X., & He, Q. (2021). Machine learning in natural and engineered water systems. *Water Research*, *205*, 117666.

Huang, R., Ma, C., Ma, J., Huangfu, X., & He, Q. (2021). Machine learning in natural and engineered water systems. *Water Research*, *205*, 117666.

Huang, R., Ma, C., Ma, J., Huangfu, X., & He, Q. (2021). Machine learning in natural and engineered water systems. *Water Research*, *205*, 117666.

Hyllestad, S., Amato, E., Nygård, K., Vold, L., & Aavitsland, P. (2021). The effectiveness of syndromic surveillance for the early detection of waterborne outbreaks: a systematic review. *BMC Infectious Diseases*, *21*, 1-12.

Iftikhar, S., Gill, S. S., Song, C., Xu, M., Aslanpour, M. S., Toosi, A. N., ... & Uhlig, S. (2023). AI-based fog and edge computing: A systematic review, taxonomy and future directions. *Internet of Things*, *21*, 100674.

Jiménez-Rodríguez, M. G., Silva-Lance, F., Parra-Arroyo, L., Medina-Salazar, D. A., Martínez-Ruiz, M., Melchor-Martínez, E. M., ... & Sosa-Hernández, J. E. (2022). Biosensors for the detection of disease outbreaks through wastewater-based epidemiology. *TrAC Trends in Analytical Chemistry*, *155*, 116585.

Khosla, N. K., Lesinski, J. M., Colombo, M., Bezinge, L., deMello, A. J., & Richards, D. A. (2022). Simplifying the complex: accessible microfluidic solutions for contemporary processes within in vitro diagnostics. *Lab on a Chip*, *22*(18), 3340-3360.

Kiyasseh, D., Zhu, T., & Clifton, D. (2020). The promise of clinical decision support systems targetting low-resource settings. *IEEE Reviews in Biomedical Engineering*, *15*, 354-371.

Kwikima, M. M. (2024). Analyzing the presence of microbial contaminants in water sourced from shallow wells within Dodoma city, Tanzania. *International Journal of Energy and Water Resources*, 1-12.

Li, Y., Wang, X., Zhao, Z., Han, S., & Liu, Z. (2020). Lagoon water quality monitoring based on digital image analysis and machine learning estimators. *Water research*, *172*, 115471.

Liu, X., Lu, D., Zhang, A., Liu, Q., & Jiang, G. (2022). Data-driven machine learning in environmental pollution: gains and problems. *Environmental science & technology*, *56*(4), 2124-2133.

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91-99.

Malabadi, R. B., Sadiya, M. R., Kolkar, K. P., & Chalannavar, R. K. (2024). Pathogenic Escherichia coli (E. coli) food borne outbreak: Detection methods and controlling measures. *Magna Scientia Advanced Research and Reviews*, *10*(1), 052-085.

Menghani, G. (2023). Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys*, *55*(12), 1-37.

Mukonza, S. S., & Chiang, J. L. (2023). Meta-Analysis of Satellite Observations for United Nations Sustainable Development Goals: Exploring the Potential of Machine Learning for Water Quality Monitoring. *Environments*, *10*(10), 170.

Ngingo, B. L., Mchome, Z. S., Bwana, V. M., Chengula, A., Mwanyika, G., Mremi, I., ... & Mboera, L. E. (2023). Socioecological systems analysis of potential factors for cholera outbreaks and assessment of health system's readiness to detect and respond in Ilemela and Nkasi districts, Tanzania. *BMC Health Services Research*, *23*(1), 1261.

Nigeria Centre for Disease Control and Prevention. (2024). *An update of cholera outbreak in Nigeria 2024*.

Okafor, N. U., Alghorani, Y., & Delaney, D. T. (2020). Improving data quality of low-cost IoT sensors in environmental monitoring networks using data fusion and machine learning approach. *ICT Express*, *6*(3), 220-228.

Omeka, M. E. (2024). Exploring the recent trends, progresses, and challenges in the application of Artificial intelligence in water quality assessment and monitoring in Nigeria: A systematic review.

Oon, Y. L., Oon, Y. S., Ayaz, M., Deng, M., Li, L., & Song, K. (2023). Waterborne pathogens detection technologies: advances, challenges, and future perspectives. *Frontiers in Microbiology*, *14*, 1286923.

Owusu, M. (2023). *Deprived Area Mapping Using a Scalable, Transferable and Open-Source Machine Learning Approach* (Master's thesis, The George Washington University).

Palermo, S. A., Maiolo, M., Brusco, A. C., Turco, M., Pirouz, B., Greco, E., ... & Piro, P. (2022). Smart technologies for water resource management: An overview. *Sensors*, *22*(16), 6225.

Polat, H., Topalcengiz, Z., & Danyluk, M. D. (2020). Prediction of Salmonella presence and absence in agricultural surface waters by artificial intelligence approaches. *Journal of Food Safety*, *40*(1), e12733.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Ranasinghe, S., & Fhogartaigh, C. N. (2021). Bacterial gastroenteritis. *Medicine*, *49*(11), 687-693.

Rebuffi, S. A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., & Mann, T. A. (2021). Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, *34*, 29935-29948.

Riccio, V., Jahangirova, G., Stocco, A., Humbatova, N., Weiss, M., & Tonella, P. (2020). Testing machine learning based systems: a systematic mapping. *Empirical Software Engineering*, *25*, 5193-5254.

RM, S. P., Maddikunta, P. K. R., Parimala, M., Koppu, S., Gadekallu, T. R., Chowdhary, C. L., & Alazab, M. (2020). An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture. *Computer Communications*, *160*, 139-149.

Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences*, *2*(7), 1308.

Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., ... & Mellit, A. (2023). A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability*, *15*(7), 5930.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 160.

Satish, N., Anmala, J., Rajitha, K., & Varma, M. R. (2024). A stacking ANN ensemble model of ML models for stream water quality prediction of Godavari River Basin, India. *Ecological Informatics*, *80*, 102500.

Sha, J., Li, X., Zhang, M., & Wang, Z. L. (2021). Comparison of forecasting models for real-time monitoring of water quality parameters based on hybrid deep learning neural networks. *Water*, *13*(11), 1547.

Shyu, H. Y., Castro, C. J., Bair, R. A., Lu, Q., & Yeh, D. H. (2023). Development of a Soft Sensor Using Machine Learning Algorithms for Predicting the Water Quality of an Onsite Wastewater Treatment System. *ACS Environmental Au*, *3*(5), 308-318.

Silva, G. M. E., Campos, D. F., Brasil, J. A. T., Tremblay, M., Mendiondo, E. M., & Ghiglieno, F. (2022). Advances in technological research for online and in situ water quality monitoring—A review. *Sustainability*, *14*(9), 5059.

Snazelle, T. T. (2020). *Field comparison of five in situ turbidity sensors* (No. 2020-1123). US Geological Survey.

Some, S., Mondal, R., Mitra, D., Jain, D., Verma, D., & Das, S. (2021). Microbial pollution of water with special reference to coliform bacteria and their nexus with environment. *Energy Nexus*, *1*, 100008.

Talnikar, M., Anmala, J., Venkateswarlu, T., & Parimi, C. (2024). Support vector machine (SVM) model development for prediction of fecal coliform of Upper Green River Watershed, Kentucky, USA. *Sustainable Water Resources Management*, *10*(3), 114.

Varoquaux, G., & Colliot, O. (2023). Evaluating machine learning models and their diagnostic value. *Machine learning for brain disorders*, 601-630.

Wang, Y., Li, S., Lin, Y., & Wang, M. (2021). Lightweight deep neural network method for water body extraction from high-resolution remote sensing images with multisensors. *Sensors*, *21*(21), 7397.

Weil, A. A., & LaRocque, R. C. (2020). Cholera and other vibrios. In *Hunter's Tropical Medicine and Emerging Infectious Diseases* (pp. 486-491). Elsevier.