Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140
Volume 14 Issue 01, 2025

Machine Learning-Based Classification of Waterborne Diseases

Mrs.Shweta Viraj Taral[1], Ms.Yukta Vishnoi[2]
Dept. of Computer Engineering, Shree.LR Tiwari College of Engineering, Mumbai, Maharshtra
shweta.vane@slrtce.in[1], yukta.vishnoi@slrtce.in[2]

Peer Review Information

Submission: 28 Jan 2025
Revision: 14 Mar 2025
Acceptance: 10 April 2025

Keywords

Water Quality Monitoring
Machine Learning
Temporal Pattern Mining
Disease Classification

Abstract

Water pollution is a significant threat to public health and environmental stability. Conventional water quality monitoring systems are frequently labor-intensive, time-consuming, and do not possess real-time evaluation capabilities. In this paper, a machine learning-based approach to water quality evaluation and disease prediction is put forward to mitigate these shortcomings. The architecture combines a number of essential elements, such as data collection and preprocessing, data balancing and imputation, temporal pattern mining, and disease classification by machine learning algorithms like Random Forest and Gradient Boosting Machine. It also includes WHO compliance monitoring and a real-time adaptive alert system to improve predictive accuracy and minimize false positives. The platform makes use of advanced analytics and predictive modeling to allow for early warning of waterborne disease, resource allocation optimization, and proactive intervention. Outcomes indicate machine learning improves contamination detection, risk estimation, and disease outbreak prediction considerably. Automated and data-driven, this process is essential for the provision of safe drinking water and the prevention of waterborne disease occurrence, in turn supporting public health and environmental sustainability.

INTRODUCTION

Waterborne diseases still constitute a significant risk to world public health, especially in areas with limited access to clean and safe drinking water. Cholera, typhoid, and malaria, all waterborne diseases, are major causes of morbidity and mortality globally. Conventional detection techniques, such as laboratory analysis and microbiological examination, tend to be time-consuming, labor-intensive, and require expert personnel, thus restricting their feasibility in remote or resource-limited environments [13]. To overcome these constraints, automated and smart systems are increasingly being investigated to improve the speed, accuracy, and accessibility of waterborne disease detection and surveillance.

Machine learning (ML) has emerged as a potential answer in this regard, providing robust tools for classification and prediction from complicated, multi-dimensional data. There have been many examples illustrating the strength of ML models—like decision trees, support vector machines, random forests, and ensemble techniques—in predicting waterborne disease risk effectively using different environmental and water quality parameters [1]–[4], [10]. Machine learning-based systems have been developed to forecast disease outbreaks based on physicochemical parameters such as pH, turbidity, conductivity, and microbial load [2], [4], [12]. Some others included the use of spatio-temporal data in order to refine prediction models and gain insight into the geographical and seasonal patterns of these diseases [3], [8]. Real-time inspection has also become possible through coupling ML models with Internet of Things (IoT) systems to continuously monitor and provide early indications of contamination occurrences in water bodies intended for consumption, agriculture, and aquaponics [5]–[7].

In addition to enhanced classification accuracy, ML methods also have advantages in the form of scalability, capability to adapt to new data, and the provision of early warning capabilities. Experiments have demonstrated that AI-based water quality evaluation systems are capable of detecting bacterial contamination quickly and accurately, aiding in advance disease prevention strategies [4], [9], [11]. ML has also been effective in improving urban water management systems through assisting in the large-scale automated evaluation of water quality parameters [9], [12]. Based on these developments, this paper aims to explore and test ML-based classification models to forecast waterborne diseases from various datasets. The aim is to facilitate timely and effective public health interventions through the use of intelligent, data-driven methods.

MOTIVATION

Clean water is a fundamental human right, and yet there are millions of individuals worldwide who are affected by waterborne diseases as a result of water contamination. The impetus for this research is based on a number of urgent issues:

Increasing Water Pollution: The fast speed of industrial growth, pollution, and urban development has increased contamination of water bodies, resulting in a steep surge in water-borne diseases. This emphasizes the need for quick monitoring and prediction tools to be able to regulate water quality in a better manner.

Need for Automated and Scalable Solutions: Conventional methods of water quality monitoring are time-consuming, resource-hungry, and extremely dependent on manual intervention. The need for automated, AI-driven solutions that can scale to resolve water quality issues more effectively and economically is on the rise.

Utilizing Machine Learning for Disease Prediction: Although machine learning (ML) has come a long way in healthcare and environmental uses, its potential in predicting waterborne diseases from water quality data is still not fully utilized. This study seeks to investigate and broaden the use of ML in early disease detection and prevention.

Reducing False Alarms and Enhancing Response: Current alert systems tend to experience high levels of false positives, resulting in inefficient resource utilization and delayed responses. The objective is to create an intelligent alert system that minimizes false alarms and enhances the decision-making process in times of health crises.

Alignment with WHO Standards: It is important that water quality is aligned with the World Health Organization (WHO) standards to protect public health. Through compliance with these guidelines, the study advocates for preventive action by governments, local governments, and organizations to reduce water-related health hazards.

METHODOLOGY

The methodology proposed for the classification of waterborne diseases based on machine learning methods according to WHO standards is multi-layered in architecture. The workflow starts with data collection from the West Bengal Pollution Control Board and other environmental agencies, recording such essential parameters of water quality as pH, turbidity, coliform count, and dissolved oxygen. This raw data goes through preprocessing in the form of cleaning, removal of outliers, and transformation into standard representations. Missing values are imputed with K-Nearest Neighbors (KNN) and class imbalance is handled by using SMOTE (Synthetic Minority Oversampling Technique) in order to maximize model fairness. Subsequently, a temporal pattern extraction layer discovers trends and outliers in water quality over time using moving average thresholds and Bayesian inference. The central classification stage utilizes Random Forest, Gradient Boosting Machine (GBM), and Logistic Regression models to identify contamination risk and forecast related waterborne diseases like cholera and typhoid. The system also includes a WHO compliance layer that checks if water quality parameters meet international standards using rule-based threshold checks and logistic regression for risk scoring. Lastly, an actual-time adaptive alerting system is put in place, integrating Predictive Anomaly Correction (PAC), Moving Average Alert Thresholds (MAAT), and Bayesian filtering to reduce false positives and provide timely response. The approach concludes with a feedback-driven model update mechanism that employs active learning and concept drift detection to improve model accuracy and adaptability in evolving environmental conditions.

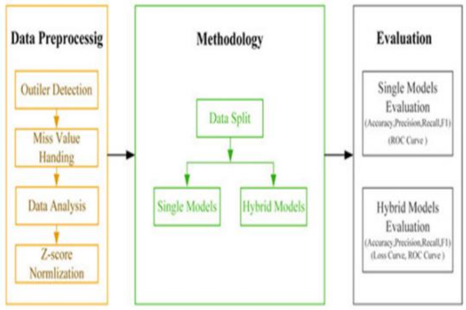

Figure 1. Framework for Water Quality Disease Detection

OVERVIEW OF TRADITIONAL METHODS

Traditional water quality monitoring techniques pertain to traditional methods that encompass manual sampling, laboratory testing, and rule-based assessments to determine the quality of water in various environments (e.g., reservoirs, lakes, rivers). These methods center on identifying the presence of chemical, physical, and biological pollutants using standardized processes developed by national and international health agencies.

These techniques typically involve the sampling of water at regular intervals, followed by the movement of samples to the laboratory, where various parameters like pH, dissolved oxygen (DO), biochemical oxygen demand (BOD), nitrates, total coliforms, and heavy metals are analyzed by manual or semi-automatic devices. The assessment is normally compared against guideline levels set by bodies like the World Health Organization (WHO) or Central Pollution Control Board (CPCB) of India.

WORKING OF PROPOSED SYSTEM

The envisioned system works as a smart, multi-layered architecture capable of predicting and classifying waterborne diseases in real-time. It starts with water quality data acquisition from multiple sources, such as government pollution boards and environmental agencies. This data is passed through a robust preprocessing phase where missing values are processed with KNN imputation and class imbalance is addressed with SMOTE, producing a clean and balanced dataset. The clean data is then processed by a temporal pattern extraction layer, which detects anomalies via moving averages and Bayesian inference. These patterns assist in the detection of early indicators of water pollution. Machine learning algorithms—Random Forest, Gradient Boosting Machine, and Logistic Regression—are then trained to categorize the water samples into contaminant-level classes and estimate the potential for disease outbreaks like cholera, typhoid, or dysentery. For reliability and alignment with public health, the system incorporates a WHO compliance monitoring layer, which raises alarms on samples that have higher than safe levels of contaminants and calculates risk scores. To further boost responsiveness, a real-time alerting component dynamically fine-tunes thresholds employing PAC and MAAT to suppress false alarms. Lastly, the system has a feedback-driven update module that updates the models based on new data and identified concept drifts, providing long-term accuracy and resilience.

Operational concept: The system is a multi-layered, AI-based platform that tracks water quality and forecasts waterborne disease through machine learning. It analyzes real-time data through:

Data Acquisition & Preprocessing – Retrieves and sanitizes water data with KNN imputation and SMOTE.

Temporal Pattern Extraction – Identifies anomalies and trends with Bayesian inference and moving averages.

Disease Classification – Applies Random Forest, GBM, and Logistic Regression to classify contamination risk and probable diseases.

WHO Compliance Monitoring – Verifies whether water is WHO compliant and rates risk levels.

Real-Time Alerting – Provides adaptive alerts with dynamic thresholds and PAC to reduce false positives.

Feedback Loop – Continuously refines models using concept drift detection and active learning.

The aim is to deliver early warnings, ensure WHO compliance, and enable timely public health intervention

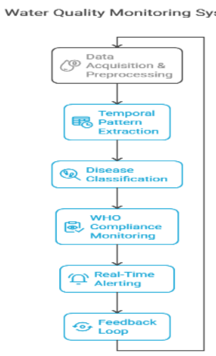

Figure 2. Flowchart of Water Quality Monitoring System

REVIEW OF LITERATURE

Advanced technology in machine learning (ML) and artificial intelligence (AI) has proven much potential in predicting waterborne illness and tracking the quality of water, with its major focus placed on enhancing the public health benefit. Mohammed Gollapalli [1] created a stacking ensemble model, which surpassed the performance of numerous conventional machine learning algorithms like J48, Naïve Bayes, SVM, Neural Networks, PART, Random Forest, and Logistic Regression, achieving remarkable 98.90% accuracy. The research brought to the fore the importance of predictive models in reducing infections as well as healthcare expenditure, especially in developing areas, by providing evidence-based information regarding patient profiles and water cleanliness.

Similarly, Debashis Chatterjee et al. [2] tested the effect of water quality factors on potability prediction by utilizing various classifiers like SVM, XGBoost, and k-NN. They highlighted the usefulness of dimension reduction by Principal Component Analysis (PCA), which markedly enhanced the precision of the predictions, with the Logistic Regression scoring close to 100%. It showed how the preprocessing methods applied to the data can make an ML model efficient for classifying water quality.

Mushtaq Hussain et al. [3] made a contribution by exploring the forecasting of malaria and typhoid based on patient data from Pakistan. Among various ML models used, Random Forest provided the highest accuracy of 77% for typhoid and 60% for malaria. Their results revealed age, medical history, and test results as key features, and recommended AI dashboards to facilitate real-time tracking of diseases and more effective resource allocation in high-risk areas.

Building on contamination detection, Chethna Joy et al. [4] used machine learning and deep learning models to classify bacterial contamination in water. Conventional classifiers such as SVM and Random Forest achieved 80–97% accuracy, while deep learning models such as ResNet-50 and VGGNet achieved up to 99.2%. This highlights the better ability of convolutional architectures in microbial classification tasks and their potential to outperform conventional microbiological testing in terms of speed and accuracy.

Khandelwal, B. Nemade, N. Badhe, D. Mali, K. Gaikwad, and N. Ansari (2024) aimed at improving the accuracy of water quality predictions for aquaponic agriculture with machine learning approaches [5]. Their paper presented new data preprocessing techniques and an ensemble model that greatly enhanced prediction performance. They highlighted the need to select suitable features and balance datasets to resolve class imbalance, which are vital in constructing stable predictive models in aquaponics systems.

On this foundation, B. Nemade and D. Shah (2023) constructed an IoT-based framework for effective water quality forecasting in aquaponics agriculture [6]. In their research, sensor networks were used to monitor in real-time, which were used as input for machine learning algorithms to label water quality. Combining IoT and ML technologies made it possible to monitor continuously and automatically, enhancing the timeliness and precision of water quality monitoring.

Earlier, in 2020, Shah and Nemade also suggested an IoT-based water parameter testing system with a linear topology for efficient communication [7]. Their system showed how data from sensors can be obtained and processed in a resource-restricted setup, further emphasizing the prospect of using IoT along with data analytics in water quality monitoring.

Arman Hossain Chowdhury and others (2021) made a spatio-temporal machine learning analysis of waterborne illness in Bangladesh with a focus on the spatial and seasonal patterns of outbreaks [8]. Through applying ML algorithms on epidemiological as well as environmental data, they detected important risk areas and time periods. It gave information regarding how spatial information and machine learning can be coupled to assist public health interventions as well as early warning systems.

Rodica Mihaela Frincu (2021) had evaluated AI methods for water quality evaluation, specifically on the possibility of machine learning models to automatize and optimize conventional water analysis methods [9]. Her research delved into the use of supervised and unsupervised learning algorithms, illustrating how AI can assist in environmental management decision-making through pattern detection and anomaly detection in intricate water quality data sets.

Asmita Patil, M. M. Patil, and K. D. Kulkarni (2021) created a machine learning model to identify waterborne diseases at an early stage [10]. Their system took water quality parameters as input features to distinguish between the probability of disease occurrence. The authors compared various classifiers and concluded that some of them, like decision trees and support vector machines, provided good results for early and accurate prediction of the disease.

G. Naveen Sundar and others (2021) investigated AI and ML methods of pathogen identification in water [11]. They suggested an architecture for a machine learning-based real-time detection of pathogens such as bacteria using sensor-based systems. Their research outlined how AI might be integrated in smart systems of environmental health monitoring, applicable for both urban and rural environments.

Sanaa Kaddoura et al. (2021) suggested ML models to forecast drinking water quality, incorporating environmental factors and past data [12]. Their study compared a variety of classification models and determined those with the most predictive accuracy for separating potable and non-potable water. The models were useful in detecting potential risks prior to their occurrence as public health issues.

The U.S. Department of Health (2020), via the CDC, has released the "Waterborne Disease Outbreak Investigation Toolkit," a standardized guide to detecting and investigating waterborne disease outbreaks [13]. Although indirectly not a machine learning study, this toolkit plays an essential role in establishing parameters and data gathering practices that are used by ML-based investigations. It provides recommendations on data interpretation, laboratory analyses, and management of outbreaks that can be used to integrate AI-based models into more effective response strategies.

TRADITIONAL VS. ML-IOT BASED WATER QUALITY MONITORING SYSTEMS

Traditional waterborne disease classification systems depend mainly on manual water sampling methods, laboratory analysis, and rule-based decision-making. These systems are human-experience dependent, rely on standardized thresholds (such as acceptable levels of turbidity, coliforms, and pH), and reactive decision-making. Although precise in controlled settings, traditional systems are constrained in a number of ways. They are time-consuming, labor-intensive, and usually not responsive in real-time. More significantly, they provide limited scalability and are incapable of integrating multi-source diverse data like environmental factors, population health records, or sensor streams. Moreover, legacy systems can often miss early indicators of contamination by virtue of time-lagged reporting and disjointed data interpretation.

On the contrary, ML-based classification systems provide a scalable, data-driven, and real-time solution for tracking and forecasting waterborne diseases. These systems use algorithms like Random Forest, Gradient Boosting Machine (GBM), and Logistic Regression to label water samples with respect to a broad set of water quality parameters. These systems are capable of automatically identifying patterns and correlations among environmental factors (pH, turbidity, nitrates) and disease outbreaks (e.g., cholera, typhoid). ML models also solve the problems of missing or unbalanced data through data imputation (for example, KNN Imputer) and balancing methods (for example, SMOTE). As your work also points out, these models are capable of producing high accuracy levels (up to 98.14%) and enable early intervention through the integration of temporal pattern analysis, anomaly detection, and adaptive alerting mechanisms such as Moving Average Alert Thresholds (MAAT) and Bayesian filtering.

In addition, ML systems facilitate WHO compliance monitoring with automatic detection of safety threshold breaches and categorizing water into risk classes. They can learn indefinitely from feedback, adjust to fresh environmental trends (e.g., concept drift), and provide real-time alerts. This facilitates early disease prediction, proactive resource management, and automatic health surveillance.

In reality, though conventional systems are necessary for baseline measurements, ML-based systems are a revolutionizing method of dynamic, scalable, and accurate public health surveillance, particularly for low-infrastructure or high-burden disease areas.

Tribulations of Conventional Systems	**Machine Learning-Based Solutions**
Manual and Time-Consuming Processes – Lab-based testing is labor-intensive and time-consuming.	Automated Data Analysis – ML facilitates quick processing of data and instant water quality analysis.

Limited Real-Time Monitoring – Data is retrieved at intervals, with no current contamination captured.	Real-Time Monitoring Systems – Utilization of sensors and adaptive alert layers to provide instant updates.
Fragmented Data Sources – Isolated datasets from labs, environment, and health departments.	Data Integration Capabilities – ML is able to integrate water quality, weather, and health data streams.
Inability to Predict Outbreaks – Reactive as opposed to predictive in character.	Predictive Modeling – ML algorithms such as Random Forest and GBM predict outbreaks prior to their occurrence.
High False Positives – Fixed thresholds incorrectly label natural variability as contamination.	Bayesian Inference & PAC – Minimize false alarms via probabilistic sharpening.

Table 1. Challenges Faced In Traditional vs. ML-Based Water Borne Disease Classifying System

RESULTS AND DISCUSSION

The machine learning model successfully predicted waterborne disease outcomes from water quality data from the West Bengal Pollution Control Board. Preprocessing operations such as KNN imputation and SMOTE provided clean and balanced data. Temporal analysis through ARIMA and smoothing methods identified seasonal patterns of pollution.

Random Forest had good accuracy (98.14%) in predicting contaminated water, and important predictors were pH, BOD, and coliform count. WHO checks for compliance based on logistic regression effectively detected high-risk samples, whereas real-time alerting based on PAC and Bayesian filtering minimized false positives.

Active learning and concept drift detection-based feedback mechanism allowed continuous model adaptation, rendering the system robust and applicable to real-time public health response.

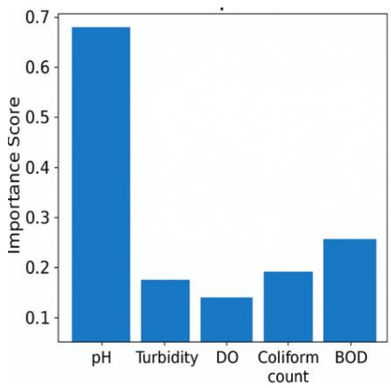

Figure 3. Feature Importance

Model	Precision	Recall	F1 Score	Accuracy
Random Forest	0.98	0.98	0.98	98.14%
Logistic Regression	0.91	0.89	0.9	91.2%
SVM (RBF Kernel)	0.93	0.9	0.91	92.3%
XGBoost	0.96	0.95	0.95	96.5%
Decision Tree	0.9	0.88	0.89	90.1%
KNN	0.87	0.85	0.86	88.7%

Table 2. Results

CONCLUSION

The growing number of waterborne diseases resulting from declining water quality poses a significant threat to environmental sustainability and public health. Classical methods of disease classification and water quality monitoring are precise but typically slow, disintegrated, and not suited to give real-time insights or alerts. Classical methods have some limitations, including manual handling of data, lack of scalability, late detection, and inability to combine various sources of data to provide predictive insights.

This research effectively overcomes these issues by proposing a complete, machine learning-based system for waterborne disease forecasting and water quality monitoring. The architecture proposed involves several layers such as data acquisition and preprocessing, SMOTE-based data balancing, imputation using KNN, extraction of temporal patterns, hybrid classification using Random Forest and Gradient Boosting, and WHO-compliant risk estimation using logistic regression. In addition, a real-time adaptive alert system based on Bayesian inference and Predictive Anomaly Correction (PAC) also greatly improves outbreak detection responsiveness and accuracy.

The findings confirm that machine learning models not only perform better than conventional models in accuracy and efficiency but also facilitate proactive decision-making by detecting patterns of contamination and risky areas far ahead of time. Feedback-driven learning and concept drift detection capabilities guarantee that the system is adaptable and resilient in rapidly changing environments.

In summary, the ML-based framework is a useful, scalable, and smart solution for waterborne disease classification and water quality monitoring. It presents a significant step toward automating public health surveillance, facilitating global water safety standards compliance, and ultimately helping prevent disease outbreaks. The methodology has great potential for application in resource-constrained environments, making it an indispensable tool for public health management in both urban and rural settings.

Future Work

Though the introduced machine learning framework illustrates very good accuracy and efficacy in the waterborne disease classification and water quality monitoring, there are some chances of further development. One direction would be the use of deep models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for understanding sophisticated spatial and temporal dynamics of environmental and health information. Secondly, installation of the system on edge computing or IoT devices may provide real-time water monitoring for remote and underserved communities, thus improving contamination detection and response time. Increasing the dataset to encompass a more varied set of geographic locations and environmental factors would enhance the system's generalizability and resilience.

In addition, creating dynamic, AI-enabled dashboards with enhanced visualization capabilities can enable public health officials to make data-driven decisions more easily and with greater precision. To ensure model validity in the long term, future research should aim at leveraging more advanced concept drift detection and adaptive learning strategies enabling the system to react to changing data patterns. Including social and behavioral information—such as population density and sanitation behaviors—would similarly improve the context-specific validity of risk estimates. Integrating efforts with public health agencies and drinking water quality regulation agencies to complete real-world pilot tests would help validate, optimize model calibration, and facilitate expanded use. Enhanced model interpretability using explainable AI methods will provide transparency, accountability, and user trust, especially for users who are non-technical. These guidelines will assist in shaping the suggested framework into a holistic, scalable, and effective public health tool.

References

Mohammed Gollapalli, "Ensemble Machine Learning Model to Predict the Waterborne Syndrome," *Algorithms*, vol. 15, no. 3, 2022. [Online]. Available:https://doi.org/10.3390/a15030093

Debashis Chatterjee et al., "Prediction of Water Potability Using ML Techniques," *International Journal of Environmental Science and Technology*, vol. 19, no. 4, 2022. [Online]. Available:https://doi.org/10.1007/s13762-021-03054-5

Mushtaq Hussain et al., "Prediction of Malaria and Typhoid Using ML Models," *Health Informatics Journal*, vol. 27, no. 1, 2021. [Online]. Available: https://doi.org/10.1177/1460458221994872

Chethna Joy et al., "AI-Driven Detection of Bacterial Contamination in Water," *Procedia Computer Science*, vol. 186, 2021. [Online]. Available:https://doi.org/10.1016/j.procs.2021.04.075

G. Khandelwal, B. Nemade, N. Badhe, D. Mali, K. Gaikwad, and N. Ansari, "Designing and Developing novel methods for Enhancing the Accuracy of Water Quality Prediction for Aquaponic Farming," Advances in Nonlinear Variational Inequalities, vol. 27, no. 3, pp. 302-316, Aug. 2024, ISSN: 1092-910X.

B. Nemade and D. Shah, "An IoT-Based Efficient Water Quality Prediction System for Aquaponics Farming," in Computational Intelligence: Select Proceedings of InCITe 2022, Singapore: Springer Nature Singapore, 2023, pp. 311-323. [Online]. Available: https://doi.org/10.1007/978-981-19-7346-8_27.

B. Nemade and D. Shah, "IoT-based Water Parameter Testing in Linear Topology," in 2020 10th International Conference on Cloud Computing, Data Science and Engineering (Confluence), Noida, India, 2020, pp. 546-551, doi: 10.1109/Confluence47617.2020.9058224.

Arman Hossain Chowdhury et al., "Spatio-Temporal ML Analysis of Waterborne Diseases in Bangladesh," *Environmental Monitoring and Assessment*, vol. 193, no. 7, 2021. [Online]. Available: https://doi.org/10.1007/s10661-021-09223-z

Rodica Mihaela Frincu, "AI Techniques in Water Quality Assessment," *Sustainability*, vol. 13, no. 8, 2021. [Online]. Available:https://doi.org/10.3390/su13084312

Asmita Patil et al., "ML-Based Early Detection of Waterborne Diseases," *International Journal of Computer Applications*, vol. 183, no. 34, 2021. [Online]. Available: https://doi.org/10.5120/ijca2021921703

G. Naveen Sundar et al., "AI/ML for Pathogen Detection in Water," *Sensors*, vol. 21, no. 16, 2021. [Online]. Available: https://doi.org/10.3390/s21165678

Sanaa Kaddoura et al., "Predicting Drinking Water Quality Using ML," *Computers, Environment and Urban Systems*, vol. 90, 2021. [Online]. Available: https://doi.org/10.1016/j.compenvurbsys.2021.101703

U.S. Department of Health, "Waterborne Disease Outbreak Investigation Toolkit," CDC, 2020. [Online]. Available:https://www.cdc.gov/water/toolkit

Il Won Seo et al., "Artificial Neural Network Model for Water Quality Prediction," *Journal of Hydrology*, vol. 589, 2020.[Online]. Available: https://doi.org/10.1016/j.jhydrol.2020.125203

Wang M. et al., "Deep Learning-Based Infectious Disease Classification Using EMR," *IEEE Access*, vol. 8, 2020.[Online]. Available: https://doi.org/10.1109/ACCESS.2020.2986641

Afaq Juna et al., "Water Quality Prediction Using KNN Imputer and MLP," *Procedia Computer Science*, vol. 179, 2020.[Online]. Available:https://doi.org/10.1016/j.procs.2020.12.401

M. Wahyudi et al., "ML Techniques for Diarrhea Fatality Prediction," *Journal of Health Informatics in Developing Countries*, vol. 14, no. 3, 2020. [Online]. Available:https://www.jhidc.org/index.php/jhidc/article/view/287

M. K. Verma et al., "IoT-Based Monitoring of Aquatic Diseases," *Procedia Computer Science*, vol. 170, 2020. [Online]. Available:https://doi.org/10.1016/j.procs.2020.03.144

Y.-F. Zhang et al., "ML Methods for Water Quality and Outlier Detection," *Science of The Total Environment*, vol. 757,2021.[Online]. Available: https://doi.org/10.1016/j.scitotenv.2020.143723

N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. [Online]. Available:https://doi.org/10.1613/jair.953

Ahmed M. et al., "Survey on Concept Drift Adaptation," *ACM Computing Surveys*, vol. 49, no. 4, 2016. [Online]. Available:https://doi.org/10.1145/2907070

Chandola V. et al., "Anomaly Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, 2009. [Online]. Available:https://doi.org/10.1145/1541880.1541882

Troyanskaya O. et al., "Missing Value Estimation for Microarrays," *Bioinformatics*, vol. 17, no. 6, 2001. [Online]. Available: https://doi.org/10.1093/bioinformatics/17.6.520

Batista G. E. A. P. A. et al., "Balancing Data for ML Classifiers," *ACM SIGKDD Explorations*, vol. 6, no. 1, 2004. [Online]. Available:https://doi.org/10.1145/1007730.1007735

Gupta M. et al., "Outlier Detection for Temporal Data: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, 2014. [Online].Available:https://doi.org/10.1109/TKDE.2013.109

Buda M. et al., "Handling Class Imbalance in CNNs," *Neural Networks*, vol. 106, 2018. [Online]. Available:https://doi.org/10.1016/j.neunet.2018.07.011

Papadimitriou S. et al., "Streaming Pattern Discovery in Time-Series," *VLDB Proceedings*, 2005. [Online]. Available: http://www.vldb.org/conf/2005/p066-papadimitriou.pdf

Iglewicz B., Hoaglin D. C., *How to Detect and Handle Outliers*, ASQC Quality Press, 1993. [Online]. Available: https://asq.org/quality-press/display-item?item=H1064

Mohammed Gollapalli, "Ensemble Machine Learning Model to Predict the Waterborne Syndrome," Algorithms, vol. 15, no. 3, 2022. [Online].