engineering proceedings

MDPI

Proceeding Paper

Performance Evaluation of Machine Learning Methods for Drinking Water Contamination Detection †

Valts Urbanovičs [1,*,†], Sergei Parshutin [2,*], Jānis Rubulis [1,*], Mārtiņš Bonders [2], Katrīna Dambeniece [1], Roberts Ozols [1], Dāvids Štēbelis [3] and Sandis Dejus [1,*]

[1] Water Systems and Biotechnology Institute, Faculty of Natural Sciences and Technology, Riga Technical University, Kipsala Street 6A, LV-1048 Riga, Latvia; katrina.dambeniece@rtu.lv (K.D.); roberts.ozols_2@rtu.lv (R.O.)
[2] Institute of Information Technology, Faculty of Computer Science, Information Technology and Energy, Riga Technical University, 10 Zunda Embankment, LV-1048 Riga, Latvia; martins.bonders@rtu.lv
[3] Waterson Technologies OÜ, Narva mnt 5, 10117 Tallinn, Estonia
* Correspondence: valts.urbanovics@rtu.lv (V.U.); sergei.parshutin@rtu.lv (S.P.); janis.rubulis@rtu.lv (J.R.); sandis.dejus@rtu.lv (S.D.)
† Presented at the 3rd International Joint Conference on Water Distribution Systems Analysis & Computing and Control for the Water Industry (WDSA/CCWI 2024), Ferrara, Italy, 1–4 July 2024.

Abstract: The aim of the study is to train a machine learning (ML) model for drinking water contamination detection and compare performance to statistical methods and existing anomaly detection solutions. A pilot drinking water supply system was made and equipped with drinking water quality sensors and a contamination dosing system. The results from this study demonstrated that using the statistical Mahalanobis distance (MD) method to predict the classification of drinking water measurements yields a 99% accuracy, 23% precision, and 28% F-score result (for wastewater contamination); however, the ML model yields a 99% accuracy, 98% precision, and a 98% F-score result. The results show that the application of ML methods can improve drinking water contamination detection speed and accuracy.

Keywords: contamination; drinking water; machine learning; neural networks

Citation: Urbanovičs, V.; Parshutin, S.; Rubulis, J.; Bonders, M.; Dambeniece, K.; Ozols, R.; Štēbelis, D.; Dejus, S. Performance Evaluation of Machine Learning Methods for Drinking Water Contamination Detection. *Eng. Proc.* **2024**, *69*, 110. https://doi.org/10.3390/engproc2024069110

Academic Editors: Stefano Alvisi, Marco Franchini, Valentina Marsili and Filippo Mazzoni

Published: 10 September 2024

1. Introduction

Safe drinking water is a cornerstone of public health and well-being, yet its quality is frequently compromised by contaminants ranging from biological pathogens to chemical pollutants [1,2]. Traditional methods for detecting contaminants in water involve time-intensive laboratory procedures that, while accurate, often fail to deliver immediate results critical for prompt remediation actions. It is estimated that 8% of the EU population is supplied by water from small waterwork companies that may struggle to provide effective drinking water quality control [3]. In a study of New Zealand water supply systems, there were at least seven cases of *E. coli* detected in drinking water that were largely dismissed due to incorrect sampling techniques [4]. Machine learning (ML) algorithms have emerged as powerful tools capable of enhancing our ability to detect and classify contaminants in drinking water with speed and precision [5]. Neural networks (NNs) are one of them, having proven themselves suitable for processing sensor data in anomaly detection.

There are several categories of drinking water contamination that must be controlled [6]:

- Biological contaminants (bacteria, viruses, etc.);
- Inorganic contaminants (heavy metals, etc.);
- Organic contaminants (phenols, pesticides, etc.);
- Emerging contaminants (microplastics, pharmaceuticals, etc.);
- Radiological contaminants.

In this study, a complex approach is taken by dosing contamination that may infiltrate a water distribution network under degraded network conditions. The used contamination may contain multiple categories of contamination. The changes in water quality parameters are marked as anomalies and used to train a ML algorithm whose precision is compared with the Mahalanobis distance (MD) method and the USEPA CANARY detection software method [7,8].

2. Materials and Methods

For generating anomalies, a pilot water distribution system was created that imitates a section of the water distribution network. The pilot system is made of 100 m of PVC pipe with a 25 mm inner diameter. The system also has drinking water quality sensors outlined in Table 1.

Table 1. List of drinking water quality sensors used in this study.

Parameter	Sensor
Flow	E+H Picomag DMA25
Pressure	E+H Cerabar PMP11
Total organic carbon + temperature	E+H Viomax CAS51D
Turbidity	E+H Turbimax CUS52D
pH	E+H Orbisint CPS11D
Oxidation–reduction potential	E+H Orbisint CPS12D
Electrical conductivity	E+H Condumax CLS21D
Flow cytometry cell counts	bNovate BactoSense

The pilot water supply system assembly principal diagram is shown in Figure 1.

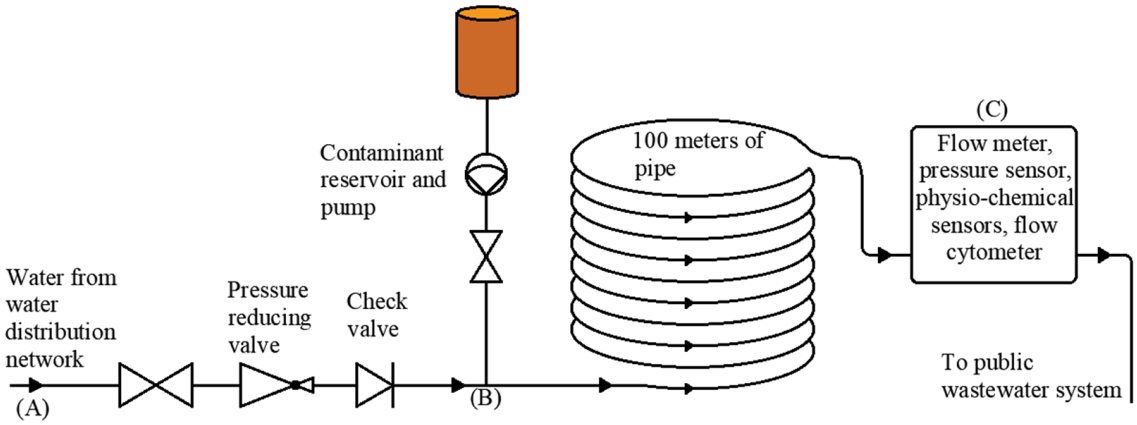

Figure 1. Pilot water supply system: (**A**) water from city water supply; (**B**) contaminant dosing junction; (**C**) sensors at the end of the pilot system.

Experiments are conducted by dosing 3 types of contamination in 3 different concentrations, repeated 3 times.

- Waste water (concentrations of 0.01%, 0.05%, and 0,5%);
- Surface water (concentrations of 0.1%, 0.5%, and 1%);
- Ground water (concentrations of 0.1%, 0.5%, and 1%).

Data are marked according to the experiment and theoretical water retention times, which are used to train an ML model. Contamination detection can be identified as a classification task, assuming each contamination type has its own footprint, differing from the drinking water. If there is no need to identify the contamination types but only to detect the fact of contamination present in the system, the task can be simplified to a binary classification task and further processed as an anomaly detection task.

NNs are applied to process the sensor data. As stated earlier, the data contained labels for readings with contamination–anomalies; thus, the supervised training model was selected to train the multilayer NNs to forecast the probability of anomaly. The NN structure used in the experiments contained 3 layers with 22,502 trainable parameters.

To distinguish critical anomalies from anomalies that were caused by some regular fluctuations in the water supply system, two different approaches were used—a simple cutoff line for minimal anomaly probability and a binomial event discriminator (BED) [9] on which the USEPA CANARY method is based.

BED uses the binomial Equation (1) to determine whether there is an Event (anomaly) currently in the system and if this Event is a Baseline Changer–critical anomaly that changes the statistical baseline parameters of the sensor readings.

$$b(r, n, p) = \frac{n!}{r!(n-r)!} p^r (1-p)^{(n-r)} = \frac{n!}{r!(n-r)!} p^r q^{(n-r)} \quad (1)$$

where

r—number of "failures" (anomalies) that occur during n trials;
n—number of repeated trials (time stamps);
p—expected probability of any one trial failing (of a reading to be an anomaly).

3. Results

The accuracy, precision recall, and F-score metrics for all three methods of anomaly detection are shown in Table 2.

Table 2. Anomaly detection metrics by anomaly detection method.

Metric	MD	BED + NN	NN
Accuracy	0.76	0.995	0.999
Precision	0.41	0.984	0.984
Recall	0.76	0.985	0.983
F-score	0.53	0.984	0.983

The BED + NN and NN anomaly detection metrics show significantly higher anomaly detection metrics than the MD method. The BED + NN and NN methods show very similar metrics.

4. Discussion

MD showed very poor performance in anomaly detection. It is possible that the MD method could show better results in tasks where specific types of anomalies must be detected. In this situation, the different types of contamination were grouped into one class that may contain differing mean values and covariance. The NN method shows good performance, which aligns with previous studies showing NNs' adequacy for data classification tasks; for example, a 95% F-score was demonstrated in another study [5]. Processing data using BED marginally improves the F-score and decreases accuracy.

5. Conclusions

Further research should be conducted by generating anomalies in different source water settings and evaluating the performance of existing anomaly detection models. Furthermore, evaluating the MD method in specific contamination detection may be needed to rule out the viability of the method for contamination detection in drinking water. Using neural networks and the binomial event discriminator for contamination detection proved to be viable methods for high-accuracy contamination detection.

Author Contributions: Conceptualization and data analysis, V.U. and S.D.; expert review, J.R.; software and data management, M.B.; conceptualization and organization of experiments, K.D. and

R.O.; project administration, D.Š.; data analysis and AI model, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the Latvian Council of Science, project "Smart Materials, Photonics, Technologies and Engineering Ecosystem" project No. VPP-EM-FOTONIKA-2022/1-0001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study. Requests to access the datasets should be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hemdan, B.A.; El-Taweel, G.E.; Goswami, P.; Pant, D.; Sevda, S. The role of biofilm in the development and dissemination of ubiquitous pathogens in drinking water distribution systems: An overview of surveillance, outbreaks, and prevention. *World J. Microbiol. Biotechnol.* **2021**, *37*, 36. [CrossRef] [PubMed]
2. Paranthaman, K.; Harrison, H. Drinking water incidents due to chemical contamination in England and Wales, 2006–2008. *J. Water Health* **2010**, *8*, 735–740. [CrossRef] [PubMed]
3. Gunnarsdottir, M.J.; Gardarsson, S.M.; Figueras, M.J.; Puigdomènech, C.; Juárez, R.; Saucedo, G.; Arnedo, M.J.; Santos, R.; Monteiro, S.; Avery, L.; et al. Water safety plan enhancements with improved drinking water quality detection techniques. *Sci. Total Environ.* **2020**, *698*, 134185. [CrossRef] [PubMed]
4. Graham, J.; Russell, K.; Gilpin, B. When the implementation of water safety plans fail: Rethinking the approach to water safety planning following a serious waterborne outbreak and implications for subsequent water sector reforms. *J. Water Health* **2023**, *21*, 1562–1571. [CrossRef] [PubMed]
5. Muharemi, F.; Logofătu, D.; Leon, F. Machine learning approaches for anomaly detection of water quality on a real-world data set. *J. Inf. Telecommun.* **2019**, *3*, 294–307. [CrossRef]
6. Sharma, S.; Bhattacharya, A. Drinking water contamination and treatment techniques. *Appl. Water Sci.* **2017**, *7*, 1043–1067. [CrossRef]
7. Dejus, S.; Nescerecka, A.; Kurcalts, G.; Juhna, T. Detection of drinking water contamination event with Mahalanobis distance method, using on-line monitoring sensors and manual measurement data. *Water Supply* **2018**, *18*, 2133–2141. [CrossRef]
8. Hart, D.B.; McKenna, S.A. *CANARY User's Manual Version 4.3.2*; Sandia National Laboratories: Albuquerque, NM, USA, 2012.
9. McKenna, S.A.; Hart, D.; Klise, K.; Cruz, V.; Wilson, M. Event Detection from Water Quality Time Series. In Proceedings of the World Environmental and Water Resources Congress 2007, Tampa, FL, USA, 15–19 May 2007. [CrossRef]