2nd International Joint Conference on Water Distribution Systems Analysis & Computing and Control in the Water Industry
Valencia (Spain), 18-22 July 2022
doi: https://doi.org/10.4995/WDSA-CCWI2022.2022.14703

MACHINE LEARNING METHODOLOGIES TO PREDICT POSSIBLE WATER QUALITY ANOMALIES AS A SUPPORT TOOL FOR ONLINE MONITORING OF ORGANIC PARAMETERS

Leonid Kadinski[1], Jonas Schuster[2], Gopinathan R. Abhijith[3], Cao Hao [4], Anissa Grieb[5], Thomas Meier[6], Pu Li[7], Mathias Ernst[8] and Avi Ostfeld F. ASCE [9]

[1] PhD Student, Faculty of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel
[2] PhD Student, Institute for Water Resources and Water Supply Hamburg University of Technology Am Schwarzenberg-Campus 3E, 21073 Hamburg, Germany
[3] Post-doctoral researcher, Faculty of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel
Schwarzenberg-Campus 3E, 21073 Hamburg, Germany
[4] PhD Student, Process Optimization Group, Institute of Automation and Systems Engineering, Technical University Ilmenau, 98693 Ilmenau, Germany
[5] Post-doctoral researcher, Institute for Water Resources and Water Supply Hamburg University of Technology Am Schwarzenberg-Campus 3E, 21073 Hamburg, Germany
[6] Head of Microbiology, Drinking Water Laboratory Hamburg Wasser, Billhorner Deich 2, 20539 Hamburg, Germany
[7] Professor, Process Optimization Group, Institute of Automation and Systems Engineering, Technical University Ilmenau, 98693 Ilmenau, Germany
[8] Professor, Institute for Water Resources and Water Supply Hamburg University of Technology Am Schwarzenberg-Campus 3E ,21073 Hamburg, Germany
[9] Professor (corresponding author), Faculty of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel

[1] kleonid@campus.technion.ac.il, [2]jonas.schuster@tuhh.de, [3]gnrabhijith@gmail.com, [4]hao.cao@tu-ilmenau.de, [5]anissa.grieb@tuhh.de, [6]thomas.meier@hamburgwasser.de, [7]pu.li@tu-ilmenau.de, [8]mathias.ernst@tuhh.de, [9] ostfeld@technion.ac.il

Abstract

Water Distribution Systems (WDSs) function to deliver high-quality water in major quantities. While standard water quality parameters are monitored at waterworks, it is still a challenge to monitor water quality in the WDS network itself. While mostly hydraulic parameters are frequently monitored and modelled in drinking water networks in Germany, the measurements of specific organic and bacteriological water quality parameter are still done offline which can take hours or even days, which might be too late to react to possible water events. This study utilizes water quality data of a Utility in Hamburg, Germany to train machine learning algorithms to predict possible anomalies in specific water quality parameters which can indicate the necessity for more thorough investigations. While a large amount of water parameters is utilized and checked for deviations from the normal distribution, the input features to train the machine learning algorithms are parameters which can be measured online like pH, temperature, total cell count of bacteria and the organic content of the water sample. A parallel study uses innovative online testing methods like fluorescence spectroscopy and flow cytometry in batch and flow experiments with the overarching goal of validating the trained algorithm to develop a wholesome online monitoring and warning system for drinking water anomalies. Various algorithms like Random Forest and Artificial Neural Networks are trained to predict whether the water samples indicate possible water quality anomalies. First results of this study show promising possibilities for a data driven online water quality prediction methodology which can help to digitalize the water sector immensely.

Keywords
Water distribution systems, contamination response, water quality monitoring, machine learning.

1 INTRODUCTION

To operate, manage and monitor water distribution systems (WDSs) is a highly complex and multidisciplinary task for water utilities. Especially monitoring water quality parameters in a continuous manner needs consideration of various boundary conditions. Specific parameters are frequently monitored at the water works because conducting measurements inside water networks is a complicated endeavour. Fixed water quality sensors and the optimized placement of them at strategically important places has shown to be highly efficient [1–4]. In Germany and in some other European countries, the drinking water in a WDS do not contain any residual disinfectants (chlorine) [5]. In these specific cases it is very important to monitor organic and bacteriological water quality parameters. Monitoring organic parameter usually requires time-intensive laboratory tests and the results are not available at once. Schuster et al (2022) [6] describe an innovative, sensitive and low-cost method how to determine organic compounds in drinking water with fluorescence sensors and flow cytometers. Recent studies have used machine learning methodologies to detect water quality anomalies and their sources by training these algorithms with specific water quality parameters to predict possible water events. Artificial and convolutional neural networks (ANN/CNN) and support vector machines (SVM) have been used to detect whether a contamination event has occurred in a water network [7,8]. Hamburg Wasser, the water utility of the city of Hamburg in Germany, provided this study with a major amount of water quality data of a time period of five years. The data from 22 sampling point in 19 water works was pre-processed and analysed with various data evaluation packages in Python. The objective is to create a support tool for water utilities as a method for exploring water quality online with e.g. flow-cytometry in real time and understanding whether additional laboratory check-ups of the water quality need to be conducted.

2 METHODOLOGY

The goal of the presented method is to develop an efficient method to detect water quality anomalies in real time by evaluating parameters that are measured online and to predict whether a drinking water sample needs follow up checks or not. After the first pre-processing and evaluation of the water quality data, the correlation of various parameters was conducted to understand for which parameter it is reasonable to train the machine learning algorithms. Although a major amount of data was provided, there was still a lack of various organic water quality parameter in the dataset. The sparse dataset was imputed with the k-Nearest Neighbor (kNN) Imputation method where the kNN algorithm is used to replace missing values. With this method, the features of the missing neighbour values are uniformly averaged or weighted according to the distance to each other [9]. While a big part of the data had to be imputed, the authors considered the imputation as sufficient for a proof of concept for the presented method. After preparing and pre-processing the data, an ANN and random forest algorithm is trained to predict water quality anomalies. The algorithms are trained with the open source Python packages Scikit-Learn [9] and TensorFlow [10]. A random forest is a classification method which belongs to the decision tree family and has various advantages regarding accuracy and efficiency compared to other models [11,12]. An artificial neural network is a machine learning model which is composed of an input layer, various hidden layers and an output layer which represents the prediction output. All of these layers consist of interconnected neurons [13–15]. A neural network is a mathematical model which predicts a specific value based on the features and data it is trained with and consist of smaller and simpler mathematical functions. Figure 1 shows a conceptual layout of an ANN where the respective data for the input and output layer of the neural network is shown. The training dataset has around 3% values which were considered as anomalies. These

are not necessarily hazardous contaminations but can also be values which vary from the normal distribution of the specific water quality parameter in the dataset.

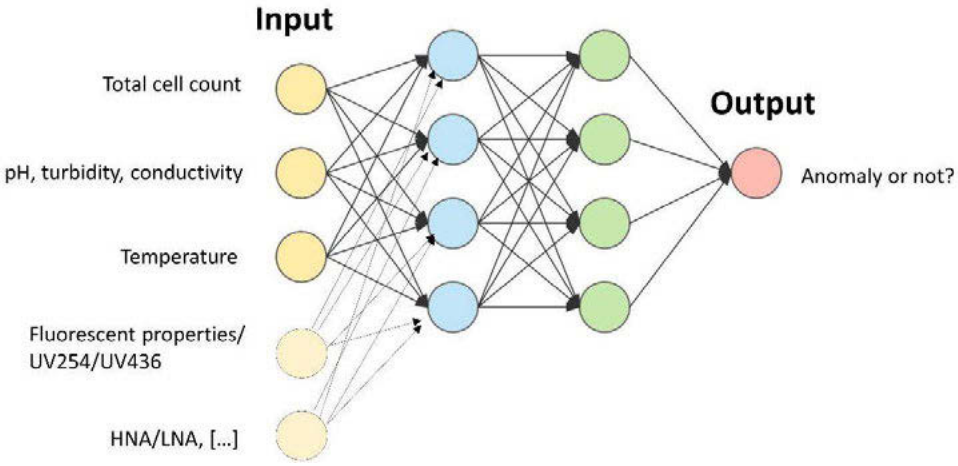

Figure 1: Conceptual layout of ANN where input features are water quality parameters that can be measured online and the output is a prediction of a possible anomaly in the water network system.

The process workflow of developing the water quality event prediction support tool follows this sequence:

- Data preparing and pre-processing
- Imputing missing data
- Training machine learning algorithm with training data set (ANN/random forest algorithm) and tuning of hyperparameter
- Testing algorithms and determining accuracy of water quality predictions with as specifically determined test data set.

3 RESULTS

The presented study introduces and presents the results of a simple case study which was conducted with the obtained water quality data training a random forest algorithm. The results which are presented were produced training a random forest algorithm to simplify the proof of concept which is meant to be shown.

Figure 2 shows the correlation of specific chosen water quality parameters in the provided dataset). The parameters chosen are all parameters which can be measured online and consequently evaluated in “real time”. They include the conductivity, total organic carbon TOC, the total cell count (TCC), all intact cells (TCCi), Turbidity, pH, UV436, temperature, O2 and high or low nucleic acid content (HNA/LNA) of the bacteria in the water. The correlation heatmap was created with the Python open source package seaborn [16].

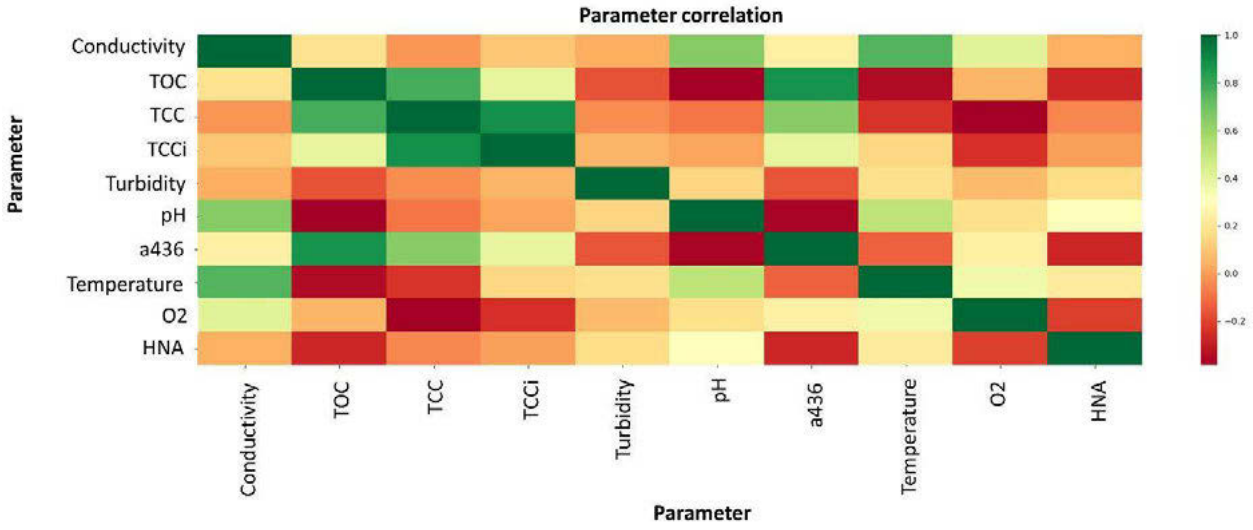

Figure 2. Correlations drawn between chosen water quality parameters

As illustrated in Figure 2, specifically the organic parameters strongly like TCC and TOC correlate with each other. However, there are correlations which might not be as clear as that, e.g., temperature with the pH or conductivity parameters of a drinking water sample. This correlation heatmap can be used to determine parameters which can be used in a sensible way for training a machine learning algorithm to give accurate predictions about the possibility of a water quality event in a WDS.

Figure 3 shows the heatmap of the confusion matrix of the water quality anomaly predictions. It was created with the Python package seaborn [16]. While the accuracy given by Scikit-Learn presents a very high value of 97%, the actual predictions need to be analysed more thoroughly. A major part of the predictions are true negatives while there have been no true positives. Which gives a false impression of the accuracy value which was put out by the software. Although it needs to be pointed out that the actual occurrences of anomalies in the dataset were comparably rare. The accuracy was tested on a test set of around three thousand samples.

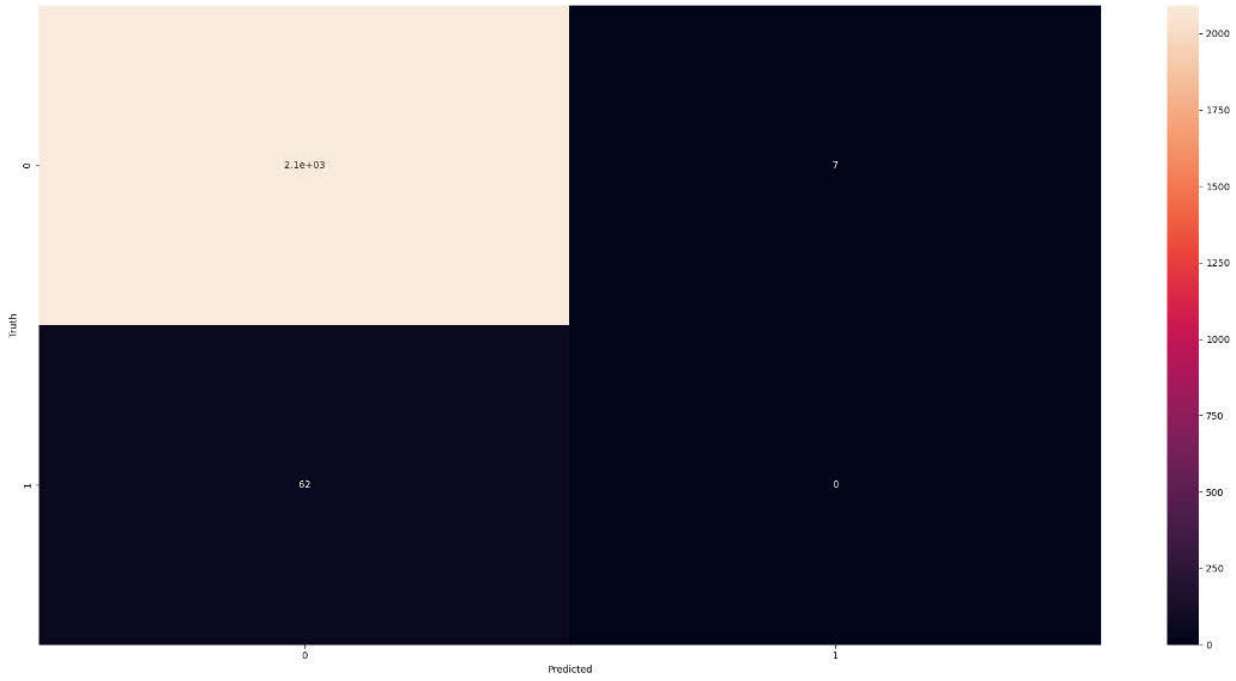

Figure 3. Confusion matrix heatmap of the water quality event predictions of the trained random forest algorithms

4 CONCLUSIONS

This study presented the proof of concept for a method to determine water quality anomalies with the online measurements of specific water quality parameters paired with a machine learning algorithm to predict a possible water event. Water quality data of the Hamburg Wasser water utility was utilized to train an ANN and a random forest algorithm, of which the latter has been presented in the results of this study. The accuracy of the random forest algorithm for predicting a water event came out to 97% and is considerably very high but it needs to be acknowledged that it came from mostly true negatives and not from true positive water event predictions. Utilizing a data driven model as a support tool for water utilities is a very promising technology and should continuously be explored with support from water utilities which have the possibility to consistently collect relevant water quality data. Future work will require more data samples so the algorithms have more bandwidth of water quality parameter that they can learn from and a more comprehensive sensitivity analysis of various machine learning algorithms to predict water quality anomalies.

5 ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support provided by the by the Israeli Ministry of Science and Technology (MOST) and the German Federal Ministry of Education and Research (BMBF) under the project numbers 3-17011 (MOST) and 02WIL1553A (BMBF).

6 REFERENCES

[1] Janke, R.; Murray, R.; Uber, J.; Taxon, T. Comparison of Physical Sampling and Real-Time Monitoring Strategies for Designing a Contamination Warning System in a Drinking Water Distribution System. 2006, 310–314.

[2] Hall, J.; Zaffiro, A.D.; Marx, R.B.; Kefauver, P.C.; Radha Krishnan, E.; Haught, R.C.; Herrmann, J.G. On-line water quality parameters as indicators of distribution system contamination. J. / Am. Water Work. Assoc. 2007, 99, doi:10.1002/j.1551-8833.2007.tb07847.x.

[3] Ostfeld, A.; Asce, M.; Salomons, E. Optimal Layout of Early Warning Detection Stations for Water Distribution Systems Security. 2004, 130, 377–385.

[4] Ostfeld, A.; Über, J.G.; Salomons, E.; Berry, J.W.; Hart, W.E.; Phillips, C.A.; Watson, J.P.; Dorini, G.; Jonkergouw, P.; Kapelan, Z.; et al. The battle of the water sensor networks (BWSN): A design challenge for engineers and algorithms. J. Water Resour. Plan. Manag. 2008, 134, 556–568, doi:10.1061/(ASCE)0733-9496(2008)134:6(556).

[5] Rosario-Ortiz, F.; Rose, J.; Speight, V.; Gunten, U. v.; Schnoor, J. How do you like your tap water? Science (80-.). 2016, 351, 912–914, doi:10.1126/science.aaf0953.

[6] Schuster, J.; Kadinski, L.; Cao, H.; Abhijith, G.R.; Grieb, A.; Li, P.; Ostfeld, A.; Asce, F.; Ernst, M. Real-time monitoring and controlling of water quality in water distribution networks based on flow cytometry and fluorescence spectroscopy. In Proceedings of the World Environmental & Water Resources Congress; 2022.

[7] Asheri Arnon, T.; Ezra, S.; Fishbain, B. Water characterization and early contamination detection in highly varying stochastic background water, based on Machine Learning methodology for processing real-time UV-Spectrophotometry. Water Res. 2019, 155, 333–342, doi:10.1016/j.watres.2019.02.027.

[8] Ashwini, C.; Singh, U.P.; Pawar, E.; Shristi Water quality monitoring using machine learning and iot. Int. J. Sci. Technol. Res. 2019, 8, 1046–1048.

[9] Pedregosa, F.; Weiss, R.; Brucher, M. Scikit-learn : Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.

[10] Martín Abadi, Ashish Agarwal, Paul Barham, E.B.; Zhifeng Chen, Craig Citro, Greg S. Corrado, A.D.; Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, I.G.; Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Y.J.; Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, M.S.; Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, J.S.; Benoit Steiner, Ilya Sutskever, Kunal Talwar, P.T.; Vincent Vanhoucke, Vijay Vasudevan, F.V.; Oriol Vinyals, Pete Warden, Martin Wattenberg, M.W.; Yuan Yu, and X.Z. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems 2015.

[11] Lee, Y.J.; Park, C.; Lee, M.L. Identification of a Contaminant Source Location in a River System Using Random Forest Models. 2018, 1–16, doi:10.3390/w10040391.

[12] Grbčić, L.; Kranjčević, L.; Družeta, S. Machine Learning and Simulation-Optimization Coupling for Water Distribution Network Contamination Source Detection. Sensors 2021, 21, 1157, doi:10.3390/s21041157.

[13] Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning; MIT Press, 2016;

[14] Fan, X.; Zhang, X.; Yu, X.B. Machine learning model and strategy for fast and accurate detection of leaks in water supply network. 2021, 6.

[15] Grbčić, L.; Lučin, I.; Kranjčević, L.; Družeta, S. Water supply network pollution source identification by random forest algorithm. J. Hydroinformatics 2020, 22, 1521–1535, doi:10.2166/HYDRO.2020.042.

[16] Waskom, M. seaborn: statistical data visualization. J. Open Source Softw. 2021, 6, 3021, doi:10.21105/joss.03021.