engineering proceedings

Proceeding Paper

Machine Learning-Based Digital Twin for Water Distribution Network Anomaly Detection and Localization [†]

Prerna Pandey [1,*], Nikolaj T. Mücke [2,3], Shashi Jain [1], Parthasarathy Ramachandran [1], Sander M. Bohté [2,4] and Cornelis W. Oosterlee [3]

[1] Department of Management Studies, Indian Institute of Science, Bangalore 560094, India; shashijain@iisc.ac.in (S.J.); parthar@iisc.ac.in (P.R.)
[2] Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands; nikolaj.mucke@cwi.nl (N.T.M.); sbohte@cwi.nl (S.M.B.)
[3] Mathematical Institute, Utrecht University, 3508 TC Utrecht, The Netherlands; c.w.oosterlee@uu.nl
[4] Swammerdam Institute of Life Sciences (SILS), University of Amsterdam, 1098 XH Amsterdam, The Netherlands
[*] Correspondence: pandey.prerna3121@gmail.com

[†] 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: This paper presents the development of a Digital Twin (DTwin) to detect and localize the leaks in water distribution networks (WDNs), using single-stage and two-stage data-driven models. In the single-stage model, we test the anomalies in the dataset using Logistic Regression and Random Forest. In the two-stage model, a linear regression model predicts pressure differences between sensor pairs in the first stage. Based on this, we compute the distribution of residuals. In the second stage, changes in the residual distribution are classified using Multinomial Logistic Regression and Random Forest models to compute possible leak locations' posterior probabilities. We have tested these models on a real-time dataset.

Keywords: digital twins (DTwins); Internet of Things; anomaly detection; AI/ML approach

check for updates

Citation: Pandey, P.; Mücke, N.T.; Jain, S.; Ramachandran, P.; Bohté, S.M.; Oosterlee, C.W. Machine Learning-Based Digital Twin for Water Distribution Network Anomaly Detection and Localization. *Eng. Proc.* **2024**, *69*, 201. https://doi.org/10.3390/engproc2024069201

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

Published: 21 October 2024

1. Introduction

Leak detection in water distribution networks (WDNs) has undergone a transformative shift with the integration of digital technologies, particularly through the deployment of IoT-enabled flow and pressure sensors. This digital approach enables real-time monitoring and analysis of pipeline conditions, enhancing the efficiency and accuracy of leak detection processes. These data are transmitted to cloud servers, where they are aggregated and analyzed using advanced algorithms for leak detection.

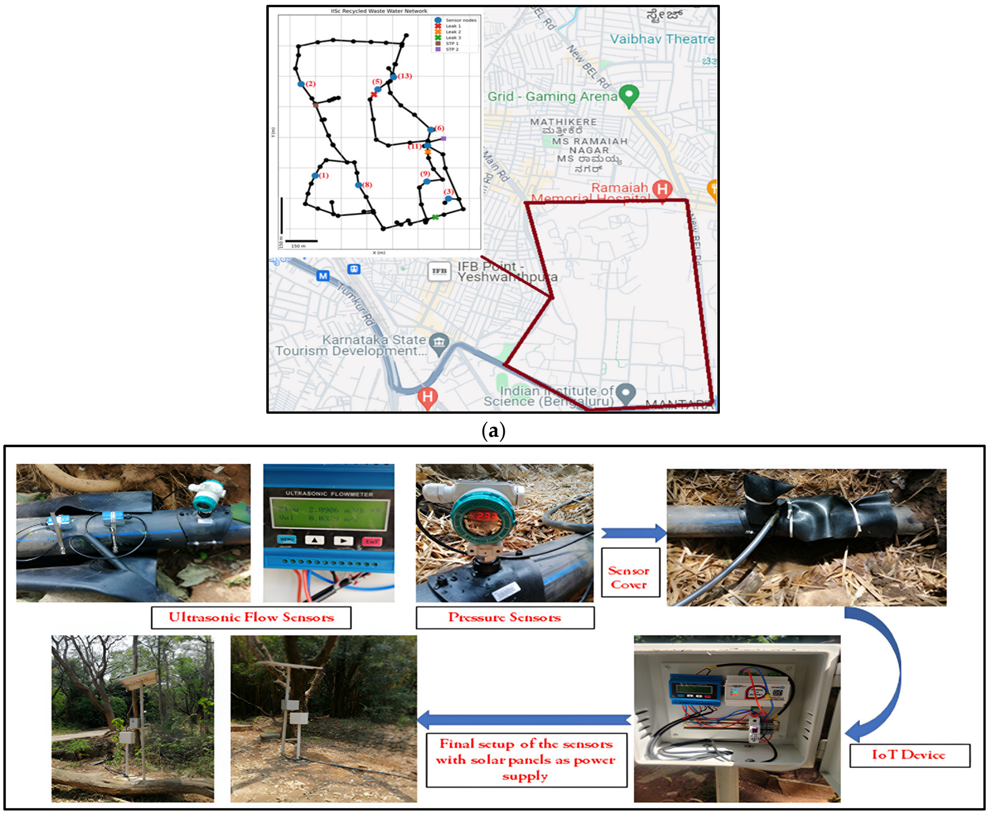
Various data-driven machine learning (ML) models for predictive applications in WDNs have been recently proposed in the literature. Ref. [1] proposed a Smart Water Grid (SWG) utilizing GIS and sensors on the water networks to monitor pressure and flow. When leakage alarms occur within the system, valve operations are initiated, resulting in a 25% reduction in leakages. Ref. [2] developed a GO2HydNet model, which updated the hydraulic model provided as an initial input. The GIS and sensor updated network was then used for leak detection using deviations between model predictions and sensor data. Ref. [3] used a two-stage process, with the first stage being the search space reduction (SSR), for instance by assuming leak occurs only in the pipe nodes and not on valves. The second stage involved a genetic algorithm-based leakage detection and localization. Ref. [4] used machine learning models on data obtained from accelerometers installed on the pipes and noise received from the surrounding environment for leak classification. They found that the deep learning model, Sqeeznet, achieved the highest accuracy. Ref. [5] used acoustic emissions with wireless noise loggers to acquire data, which were then analyzed using ML classification algorithms to detect patterns and classify leaks or no-leak cases. SVM, ANN, and deep learning (DL) models show higher accuracy, particularly with unlabeled datasets. Ref. [6] used generative deep learning to develop a probabilistic surrogate model for the WDN, replacing hydraulic equations and accounting for network uncertainties. Bayesian inference combined sensor data with the surrogate model output, providing a posterior distribution of potential leak locations and estimating leak location and uncertainty.

Ref. [7] used a two-stage model for leak detection and classification. In the first stage, the relationship between flow rate and pressure head among pairs of nodes in the network was learned. The second stage involved simultaneously using the prediction errors between these pairs of nodes to classify leaks and determine the leak location. In the extension of the previous work, we proposed a single-stage model approach based on Logistic Regression and Random Forest, where it directly takes the sensor data as an input for classification.

In this study, we created a DTwin model with a focus on real-time data collection using IoT-enabled pressure and flow sensors from the treated wastewater supply network at the Indian Institute of Science (IISc), Bangalore. Three controlled leaks were introduced in different pipe segments to test the developed leak localization model.

2. Network and Dataset

This study was carried out on the campus of IISc Bangalore and treated a wastewater supply line that was installed in the year 2009, as is shown in Figure 1a. The recycled wastewater supply line is 8.974 km long. The network comprises 90 nodes, 93 pipes, and 2 reservoirs (source was the wastewater treatment plant). The pipes exhibited varying diameters, ranging from 63 mm to 160 mm, lengths between 5 m and 245 m, and a roughness of 85. The network provided intermittent supply: the supply hours of the network were from 6 AM to 11 AM and 5 PM to 7 PM.


IISc Recycled Waste Water Network
Sensor nodes
Leak 1
Leak 2
Leak 3
STP 1
STP 2
(1)
(2)
(3)
(5)
(6)
(8)
(9)
(11)
(13)
150 m
Vaibhav Theatre
Grid - Gaming Arena
MATHIKERE
MS RAMAIAH NAGAR
Ramaiah Memorial Hospital
New BEL Rd
IFB Point - Yeshwanthpura
Tumkur Rd
Karnataka State Tourism Development...
Indian Institute of Science (Bengaluru)
Ultrasonic Flow Sensors
Pressure Sensors
Sensor Cover
IoT Device
Final setup of the sensors with solar panels as power supply

(**a**)

(**b**)

Figure 1. Schematic of DTwin setup; (**a**) layout of IISc network; (**b**) a complete setup for DTwin.

2.1. Sensors Placement

The pressure and flow sensors were placed at nine different locations within the network as shown in Figure 1a, with sensors marked by blue circles (labeled identically to those in the field). These sensors continuously record real-time data throughout the day.

2.2. Data Collection

The sensors were linked to an IoT device. The IoT device undertook the responsibility of transmitting data using the MQTT protocol. The MQTT broker, that runs on the server, established communication between the sensors and the SCADA software, where the data were stored every 5 min. The complete setup is as shown in Figure 1b. The data were also stored for every minute in a PSQL database.

2.3. Leak Location

For experimental purposes, a total of three leak locations were selected, as marked by "**x**" in Figure 1a. Three controlled leaks were induced in separate pipe segments using a tap within the network. Sensor data collected with all taps closed are labeled as "no-leak", while data with one of three taps open is labeled as leak location 1, 2, or 3, representing which tap was opened during data collection. Only one tap was opened at a time, simulating an additional unaccounted-for demand point.

2.4. Use of Machine Learning Models

To detect and localize leaks in the network, we used four supervised machine learning models. These models were grouped into two: single-stage and two-stage models. In the single-stage approach, we used **Logistic Regression** for binary classification and the **Random Forest** for classifying leaks directly.

The two-stage model comprised the **Linear Regression—Multinomial Logistic Regression Model** and the **Linear Regression—Random Forest Model**. In the first stage, we fitted linear regression models for each sensor pair using subsets of data from the no-leak scenario to predict Δhi,j for no leaks. The difference between the mean prediction error for a period and the historically sampled mean prediction error for the no-leak dataset served as an input for the second stage [7].

3. Results and Discussion

The real-time data observed from all nine flow and pressure sensors installed in the network were recorded for every minute. They were then divided into training (70%) and testing (30%) sets for analysis purposes.

3.1. Results Using Single-Stage Model

The confusion matrix obtained by single-stage models showed that the Logistic Regression and Random Forest algorithms alone can achieve an accuracy of 100%.

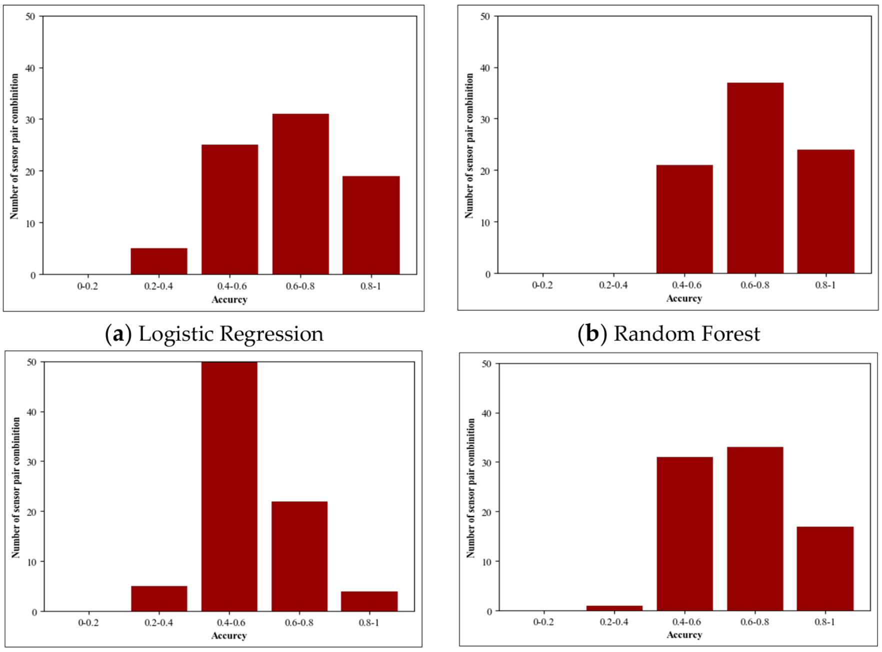
This shows that the method worked decently with the definite amount of information available for leak location.

3.2. Results Using Two-Stage Model

The confusion matrix obtained by the two-stage models, showed that the two-stage approach combining Linear and Multinomial Logistic Regression achieves 100% accuracy, while the Linear and Random Forest approach exhibits some misclassification, resulting in a compromised accuracy of 77.5%.

3.3. Comparison of Single-Stage and Two-Stage Models When a Limited Number of Sensors Are Utilized

We compared models using a subset of nine available sensors. Tests were conducted with all possible combinations of three sensors among the nine, and the results are depicted in Figure 2. Among 82 combinations, Random Forest and Logistic Regression achieved the best accuracy, with over 70% in 57 and 43 sensor combinations. In contrast, Linear Regression and Logistic Regression, and Linear Regression and Random Forest, achieved accuracy above 70% in 23 and 44 sensor combinations, respectively.



(**a**) Logistic Regression (**b**) Random Forest

(**c**) Regression + Logistic Regression (**d**) Regression + Random Forest

Figure 2. Classification of leak using single—stage and two—stage models with randomly chosen placement of 3 sets of sensors.

4. Conclusions

We have developed a DTwin model for the real-time IISc Bangalore wastewater supply network. Nine IoT-enabled flow and pressure sensors collect data every minute. We introduced three controlled leaks in different pipe segments using a tap. The data were then used for leak detection and localization with two models: single-stage, using Logistic Regression or Random Forest directly, and two-stage, using Regression to learn sensor pair relationships before classification with Logistic Regression or Random Forest.

When all nine sensor data are used, all models show high accuracy in leak prediction and localization. However, using only three pairs of sensors requires careful consideration of both model choice and sensor placement for higher accuracy. Among the models, the single-stage Random Forest was robust regardless of sensor placement. The best accuracy for all models was achieved with the most scattered sensor combination (locations 1, 2, and 9).

Author Contributions: P.P. and N.T.M.; methodology, N.T.M.; software, S.J.; validation, P.P., N.T.M. and S.J.; formal analysis, P.P. and N.T.M.; investigation, S.J.; resources, N.T.M.; data curation, P.P.; writing—original draft preparation, P.P., N.T.M. and S.J.; writing—review and editing, P.P., N.T.M. and S.J.; visualization, S.J., P.R., S.M.B. and C.W.O.; supervision, S.J. and C.W.O.; project administration, P.P., N.T.M., S.J., P.R., S.M.B. and C.W.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data can be made available on Github repository.

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

References

1. Ramos, H.M.; Morani, M.C.; Carravetta, A.; Fecarrotta, O.; Adeyeye, K.; López-Jiménez, P.A.; Pérez-Sánchez, M. New challenges towards smart systems' efficiency by digital twin in water distribution networks. *Water* **2022**, *14*, 1304–1321. [CrossRef]
2. Conejos Fuertes, P.; Martínez Alzamora, F.; Hervás Carot, M.; Alonso Campos, J. Building and exploiting a Digital Twin for the management of drinking water distribution networks. *Urban Water J.* **2020**, *17*, 704–713. [CrossRef]
3. Sophocleous, S.; Savić, D.; Kapelan, Z. Leak localization in a real water distribution network based on search-space reduction. *J. Water Resour. Plan. Manag.* **2019**, *145*, 1–15. [CrossRef]
4. Yu, T.; Chen, X.; Yan, W.; Xu, Z.; Ye, M. Leak detection in water distribution systems by classifying vibration signals. *Mech. Syst. Signal Process.* **2023**, *185*, 1–19. [CrossRef]
5. Fares, A.; Tijani, I.; Rui, Z.; Zayed, T. Leak detection in real water distribution networks based on acoustic emission and machine learning. *Environ. Technol.* **2023**, *44*, 3850–3866. [CrossRef] [PubMed]
6. Mücke, N.T.; Pandey, P.; Jain, S.; Bohté, S.M.; Oosterlee, C.W. A Probabilistic Digital Twin for Leak Localization in Water Distribution Networks Using Generative Deep Learning. *Sensors* **2023**, *23*, 6179–6198. [CrossRef] [PubMed]
7. Tyagi, V.; Pandey, P.; Jain, S.; Ramachandran, P. A two-stage model for data-driven leakage detection and localization in water distribution networks. *Water* **2023**, *15*, 2710–2725. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.