Leak localization in Water Distribution Networks By Directly Fitting the Learning Parameters of a Gaussian Naive Bayes Classifier

1st Ganjour Mazaev
IDLab
Ghent University - imec
Ghent, Belgium
ganjour.mazaev@ugent.be

2nd Michael Weyns
IDLab
Ghent University - imec
Ghent, Belgium
michael.weyns@ugent.be

3rd Filip Vancoillie
De Watergroep
Brussels, Belgium
filip.vancoillie@dewatergroep.be

4th Guido Vaes
Hydroscan
Leuven, Belgium
guido.vaes@hydroscan.be

5th Femke Ongenae
IDLab
Ghent University - imec
Ghent, Belgium
femke.ongenae@ugent.be

6th Sofie Van Hoecke
IDLab
Ghent University - imec
Ghent, Belgium
sofie.vanhoecke@ugent.be

***Abstract*—Water supply companies around the globe are struggling to meet the needs of an ever-increasing population, while climate change contributes to more drought. At the same time, up to 30% of the total amount of treated drinking water in the water supply system is lost due to leaks. An important strategy to reduce leak losses is using hydraulic modeling to localize leaks in a expert-driven manner. In this paper, we present a hybrid leak localization approach combining both hydraulic modeling and machine learning-based classification. A Gaussian Naive Bayes classifier is trained to localize leaks based on simulated pressures and historical pressure measurements. The simulated pressures are obtained using a hydraulic model of the water supply system. In our methodology, learned parameters of the classifier are inferred directly from processing the simulated and measured pressures, without the need for explicit training. We demonstrate the effectiveness of our leak localization approach by using real leak experiments, achieved by opening hydrants at different locations in an operational water supply system. State-of-the-art results are achieved, similar to an approach where explicit training is still needed.**

***Index Terms*—water distribution network, leak localization, machine learning, predictive maintenance, hydraulics**

I. Introduction

Leakages in water distribution systems (WDNs) are a major problem for water providers and their customers all around the world. Up to 30% of the total amount of treated drinking water is lost to leaks, resulting in substantial economic, societal and environmental losses [1]. Leakages result in inefficient distribution of energy within the network, wasting energy necessary for pumping the water towards customers. Leakages exacerbate water scarcity problems. Leaks also impact water quality since low pressure conditions are introduced in the WDN, potentially leading to infected water. Furthermore, leakages may lead to substantial damages caused by running water [2]. It is therefore crucial to prevent, identify, and stop water leakage in WDNs.

The most common approach for localizing water leaks in WDNs is to carry out surveys using portable measurement tools [3]. These tools detect vibrations caused by water leaking from pressurized pipelines. Although the effectiveness of this approach has been thoroughly established, it still has shortcomings. Most notably, it requires a lot of manual labor, and its accuracy depends on the expertise of the operator handling the measurement tools [4].

An alternative approach to leak localization is to use software-based methods, which eliminate the need for large-scale, manual surveys. Software-based methods rely on computational algorithms and automated data-analysis for finding anomalies in hydraulic time-series data patterns. Sensors in the field are used to monitor WDN parameters such as pressure, mass flow-rate or temperature [5]. Software-based methods can be divided into three categories: model-based approaches, data-driven approaches and hybrid approaches.

In model-based approaches, an analytic model of the WDN is built based on hydraulic laws describing its operation. This model is then used in fault detection and isolation techniques [6]. A main disadvantage of these approaches is that the localization performance is directly tied to the accuracy of the hydraulic model. Some examples of this approach have been published by Blesa *et al.* [7] and Perez *et al.* [8].

In data-driven approaches, leaks are localized based on the

The work of G. Mazaev and M. Weyns was supported by Research Foundation - Flanders (FWO) under strategic basis research doctoral grants (1S88020N, 1SD8821N). Part of this work was supported through the SmartWaterGrid project, an imec.icon research project funded by imec and Agentschap Innoveren & Ondernemen.

978-1-6654-8045-1/22/$31.00 ©2022 IEEE

application of statistical and machine learning techniques to historical data. This strategy has gained popularity in recent years, due to the increasing amount of data collected in WDNs [6]. A main drawback using purely data-driven approaches for localizing leaks, is that usually the amount of data does not suffice to represent different leak scenarios. Recent examples can be found in Zhou *et al.* [9] and Navarro *et al.* [10].

In hybrid approaches, model-based and data-driven techniques are combined to integrate their strengths into one single strategy [11]. Here, data-driven approaches (such as statistical process control or machine learning classification) are applied to data generated by a hydraulic model of the WDN [12], [13].

In this paper, we propose a hybrid leak localization approach combining hydraulic modeling and data-driven classification. Pressure measurements of various leak scenarios are simulated using a hydraulic model of the WDN. In combination with historical pressure measurements in leak-free conditions, these simulated pressures are processed into features to be used for leak location classification. The classification is performed by a Gaussian Naive Bayes (GNB) classifier. Whereas state-of-the-art data-driven and or hybrid approaches require an explicit training step, a compelling aspect of our approach is that the classifier does not actually need to be trained using the processed features. Rather than estimating the learned parameters from these features, the GNB parameters can be fitted directly using the processed pressure time-series. The leak localization performance of the presented classifier is tested using real leak experiments in an operational WDN and compared to a state-of-the-art approach that needs explicit training.

The rest of this article is organized as follows. In Section 2, the data and methodology is described. In Section 3, the results of the presented leak localization methodology are presented and discussed. Section 4 draws the conclusion of this article.

II. Methods

A. Data

In this work, we evaluate our leak localization methodology in BK-Town, a WDN managed by De Watergroep in Belgium [14]. Its topological structure is shown in Fig. 1. Hydrants are considered as potential leak locations, consisting of 360 locations in total. The hydrants of the network are visualized as nodes in the graph. Pipes are represented by edges in the graph.

A total of 9 leak experiments were performed in the WDN, which we each aim to localize. Leaks were emulated by opening hydrants for a period of 20 minutes. The resulting leaks are representative for real in-field leaks. During each leak experiment, water is lost through the hydrant at approximately 10 m^3/h. Pressure data was collected using the 19 pressure sensors shown in Fig. 1. The pressure measurement time-series are sampled at 5-minute intervals.

B. Overview of the leak localization approach

An overview of our leak localization methodology is given in Fig. 2. We aim to illustrate how a trained leak classifier

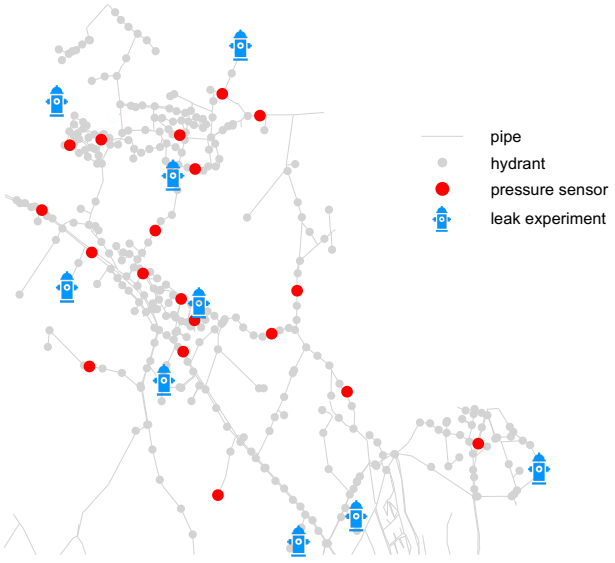

Fig. 1. Layout of the WDN, showing the location of pipes, hydrants and pressure sensors. Leaks are induced in the network by opening hydrants.

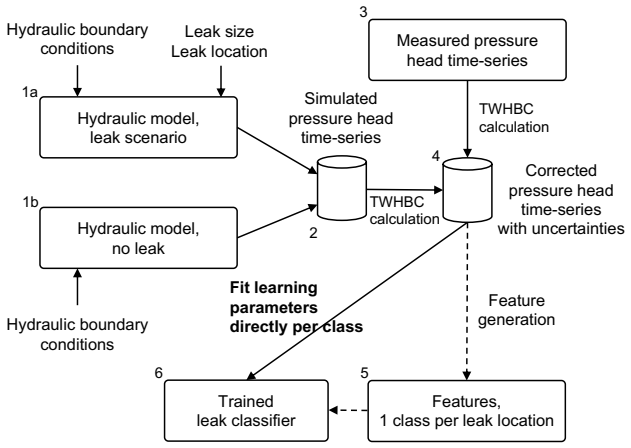

Fig. 2. Overview of the leak localization methodology.

is obtained, starting from a hydraulic WDN model and pressure measurements, without needing an explicit training step. The whole methodology is now described from a high-level perspective.

First, pressure time-series of different leaks scenarios are generated in every hydrant by simulating the WDN using a hydraulic model in leaky (1a) and leak-free conditions (1b) (see Section II-C), resulting in simulated pressure head time-series (2). A *Time-Windowed Head Bias Correction* (TWHBC) is then applied (see [14]) to these time-series, by comparing them with historical pressure measurements (3). The reason for using the TWHBC, is that bias corrections and uncertainties are added to the simulated pressures. These corrected pressures (4) are a more faithful representation of the real system, since sensor noise and hydraulic modeling errors are accounted for.

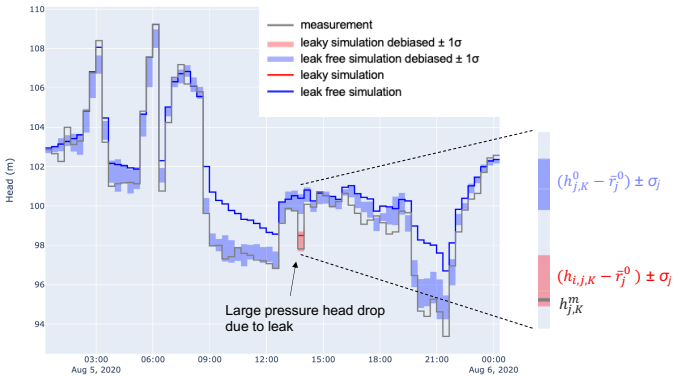

Fig. 3. Visualization of the TWHBC for one pressure sensor.

These time-series are then converted into features (5) (see Section II-D), where each leak location is represented by a class. This dataset can then be used to train a classification model (6). However, we show that it is possible to omit generating the features. Instead, the learning parameters of a Gaussian Naive Bayes classifier can be found directly by making use of the TWHBC calculations. We describe the methodology in more detail in the following sections.

C. Pressure head time-series

As is common in hydraulic WDN modeling, we make use of pressure head values (in meter units) instead of pressures (bar units). Pressure data is converted to head values by multiplication of a conversion factor, i.e., 10.1974 [14]. The elevation of the pressure measurement location is then added to obtain the final pressure head value.

The WNTR package [15] was used to obtain the simulated pressure heads of the BK-Town network. The Hazen–Williams equation is used to calculate the pipe friction factors [16]. A leak scenario in every hydrant of the WDN is simulated, by adding a demand of 10 m^3/h to the hydrant considered. As visualized in Fig. 1, there are 360 hydrants in total, with 19 pressure sensor locations. As a result, 360 sets of 19 pressure head time-series are simulated, each set simulating a leak in one hydrant as the leak location.

D. From TWHBC to training a GNB classifier

In general, measured physical properties of a physical system never correspond perfectly with their simulated counterparts obtained from a mathematical model representing the system. As for a WDN, measured pressure head values do not correspond perfectly to their simulated values obtained from a WNTR simulation, due to hydraulic modeling errors and pressure sensor noise.

Measured head values for one pressure sensor are shown as a grey line in Fig. 3. More specifically, the average head value for every interval of 20 minutes (e.g., from 12:00 to 12:20) is shown. The head values obtained from the WNTR simulation are shown as a blue line. As can be seen, the simulation can differ considerably from the measurement, depending on the time of day, since the simulation is an imperfect representation of the real system.

To correct for this difference, the difference is quantified using the 11 workdays before the day of the measurement. For example, when considering August 5, we start to calculate this difference on July 21. For example, if the interval from 12:00 to 12:20 is considered, the difference for this interval is calculated 11 times starting from July 21. The mean and standard deviation of these values are then used to correct for the difference between the current head simulation and measurement. The corrected head simulation is shown as a $\pm 1\sigma$ (i.e., the standard deviation) interval in blue in Fig. 3 for every 20-minute interval of August 5. We name this correction, including the uncertainty interval, the *Time-Windowed Head Bias Correction*.

For the interval from 13:40 to 14:00, we note that the real measured head value drops considerably. This drop occurs due to a nearby leak experiment of size 10 m^3/h happening during that interval. For this reason, we also show the head values for a leak simulation of size 10 m^3/h at this hydrant, after application of the TWHBC, as a $\pm 1\sigma$ interval in red. Since this simulation takes the head drop resulting from the leak into account, it conforms better with the measured head value.

1) Pressure head residuals: We now formally define our leak localization methodology. To avoid too many indices in the mathematical notation, we assume a given time-window (e.g. 13:40 to 14:00), without explicitly notating this window. We assume N possible leak locations in the WDN (i.e., 360 in our case), and M pressure sensor locations (i.e., 19). A leak location in node i is noted as l_i $(i = 1, ..., N)$, and a pressure sensor by j $(j = 1, ..., M)$. The day is indicated by k, with $k = K$ (i.e., 12) corresponding to the day for which the TWHBC is computed. For sensor j and day k, the following head values and residuals can then be defined:

- $h^m_{j,k}$ = the measured head.
- $h^0_{j,k}$ = the leak free simulated head.
- $h_{i,j,k}$ = simulated head for leak location l_i.
- $r^0_{j,k} = h^0_{j,k} - h^m_{j,k}$ = the head residual corresponding to the leak free simulation.

The means of the TWHBC are computed by averaging $r^0_{j,k}$ over the days preceding $k = K$:

$$\bar{r}^0_j = \frac{1}{K-1} \sum_{k=1}^{K-1} r^0_{j,k}. \tag{1}$$

Similarly, we obtain the standard deviations of the TWHBC:

$$\sigma_j = \sqrt{\frac{1}{K-2} \sum_{k=1}^{K-1} (r^0_{j,k} - \bar{r}^0_j)^2}. \tag{2}$$

The debiased leak free simulated head on day K is then given by $h^0_{j,K} - \bar{r}^0_j$. A debiased leaky simulated head is given by $h_{i,j,K} - \bar{r}^0_j$. Both share the same uncertainty of σ_j. For example, we consider the averaged head values as shown in Fig. 3, for the time interval from 13:40 to 14:00. The grey

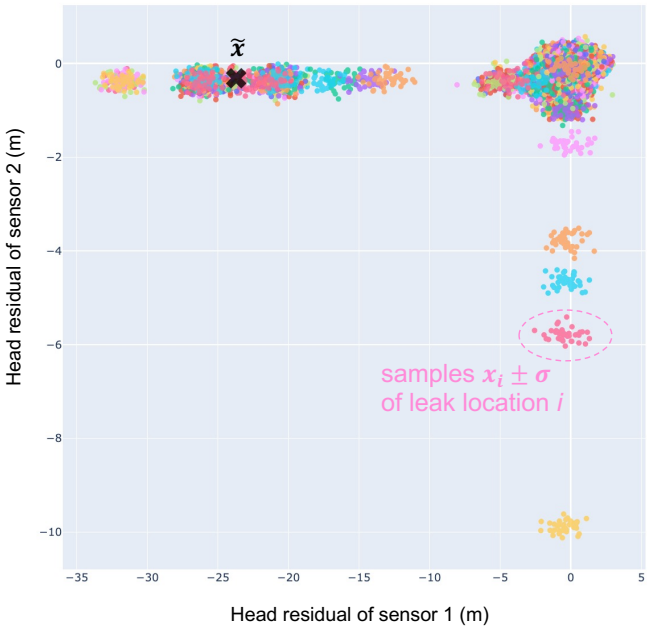

Fig. 4. Head residuals for 2 pressure sensors generated for the leak locations.

line corresponds with $h^m_{j,K}$, with j the index corresponding to the pressure sensor in the figure. $(h^0_{j,K} - \bar{r}^0_j) \pm \sigma_j$ corresponds to the leak-free simulation. $(h_{i,j,K} - \bar{r}^0_j) \pm \sigma_j$ corresponds to the leak simulation, with i indicating the hydrant leak location which was simulated.

We now define the pressure head residuals on which classification is performed. Residuals for the leaky simulations are written as

$$x_{i,j} = h^0_{j,K} - h_{i,j,K}. \tag{3}$$

Note that $\bar{r}^0_j$ does not appear in this expression, as both $h^0_{j,K}$ and $h_{i,j,K}$ are debiased with this value (as shown in Fig. 3). These residuals can be summarized in a vector over all pressure sensors $(j = 1, ..., M)$:

$$\mathbf{x}_i = \mathbf{h}^0_K - \mathbf{h}_{i,K}. \tag{4}$$

Residuals for the measured head values are written as

$$\tilde{x}_j = h^0_{j,K} - \bar{r}^0_j - h^m_{j,K}. \tag{5}$$

The corresponding residual vector is then written as:

$$\tilde{\mathbf{x}} = \mathbf{h}^0_K - \bar{\mathbf{r}}^0 - \mathbf{h}^m_K. \tag{6}$$

In Fig. 4, a visualization of the head residuals is given for 2 pressure sensors, enabling a visualization in 2D. Head residuals are shown for every leak candidate, using one colour per leak candidate. Instances of simulated residual vectors are obtained by randomly sampling the Gaussians defined by the standard deviations σ_j, and means $\mathbf{x}_i$. A leak location l_i is thus characterized by a cloud of head residual vectors typical for that leak. We also show an example of measured pressure head values $\tilde{\mathbf{x}}$ in this feature space, indicated by the black cross.

2) Training the GNB classifier: Examples of hybrid leak localization methodologies where features are constructed from pressure head residuals can be found in earlier work: in [12], [17] the uncertainties per class are caused by randomly varying demands at customer nodes and adding artificial pressure sensor noise; in [18] a similar approach is used, where pressure residuals are transformed into a cosine space. In these approaches features need to be generated in order to train a ML classifier (as exemplified in Fig. 4). We will now describe how a GNB classifier can be trained without needing to generate these features.

In the Naive Bayes assumption, a class conditional density of the following type is used:

$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{d=1}^{D} p(x_d|y = c, \boldsymbol{\theta}_{dc}) \tag{7}$$

with θ_{dc} the parameters for the class conditional densities of class c and feature d. The posterior over the class labels is then given by

$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) = \frac{p(y = c|\boldsymbol{\pi}) \prod_{d=1}^{D} p(x_d|y = c, \boldsymbol{\theta}_{dc})}{\sum_{c'} p(y = c'|\boldsymbol{\pi}) \prod_{d=1}^{D} p(x_d|y = c', \boldsymbol{\theta}_{dc'})} \tag{8}$$

where π_c is the prior probability of class c. For real-valued features x_d, univariate Gaussian distribution can be used for the class conditional densities:

$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{d=1}^{D} \mathcal{N}(x_d|\mu_{dc}, \sigma^2_{dc}), \tag{9}$$

where μ_{dc} and σ_{dc} define the Gaussian distribution of feature d with class label c [19]. Thus, a GNB classifier is a generative classifier with each class being modeled as a multivariate Gaussian with a diagonal covariance matrix. The learning parameters of the classifier are usually fit using maximum likelihood estimation.

Through the calculation of the TWHBC however, these learning parameters are already available. The means correspond with pressure residuals $\mathbf{x}_i$ in Eq. (4) and the standard deviations correspond with σ_j in Eq. (2). Hence, these parameters can be inserted directly into GNB classifier, without generating the features first.

The resulting GNB classification model is trained on leak simulations, and is used to classify real pressure head measurements. We also introduce a form of regularization, so that the classifier generalizes its predictions to real measurements ([14], [17]). To do so, we introduce an extra hyperparameter k to the class conditional densities of Eq. (9), resulting in Eq. (10). k is a scaling factor that increases the variance of each univariate Gaussian with a constant factor:

$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{d=1}^{D} \mathcal{N}(x_d|\mu_{dc}, k\sigma^2_{dc}) \tag{10}$$

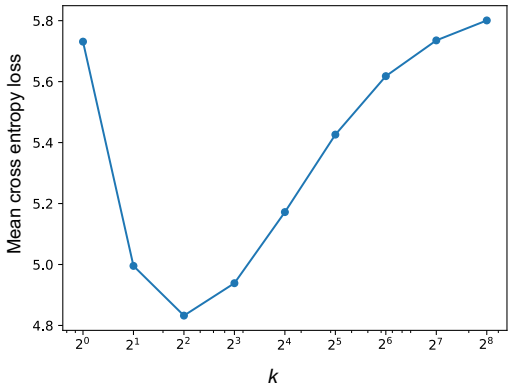

Fig. 5. Mean cross entropy over all leak experiments for each hyperparameter k.

3) Evaluation: As here only a small number (9 samples) of in-field leak experiments can be used as test data, we evaluate and report the average cross-entropy loss calculated over all leak experiments per hyperparameter k in Eq. (10). A map of the leak probability predictions for the best hyperparameter is then visualized. Results are compared to the results published in [14], where a state-of-the-art Elastic-Net logistic regression model was presented to classify the leaks in the BK-Town network.

III. Results and Discussion

The mean cross entropy over all leak experiments, computed over different values for hyperparameter k of Eq. (10) is shown in Fig. 5. As can be seen, the optimal value is reached for $k = 4$, resulting in a mean cross entropy of 4.82. This value is slightly higher than the optimal value equal to 4.80 obtained in [14], however no explicit training step was required. Predicted leak probabilities for both the GNB and Elastic-Net logistic regression models of two leak experiments are shown in Figures 6 and 7 for $k = 4$. Higher leak probabilities are shown in a darker color. We observe that the true leak locations are located in the most probable leak regions predicted by the GNB model. The results are very similar compared to the logistic regression model, except that the use of our GNB methodology results in slightly smoother changes in the leak probabilities towards the true leak location. Hence, it is interesting to observe that an almost similar leak localization performance can be achieved, without the need to actually generate the classification features and train a model on those features, showing the potential of the presented GNB approach for leak localization.

IV. Conclusion

The proposed hybrid leak localization methodology in this work uses a combination of hydraulic and data-driven modeling, applied to pressure sensor measurements in a WDN. We have shown the effectiveness of our methodology in an operational WDN, and compared its leak localization performance with a state-of-the-art Elastic-Net logistic regression model trained on pressure head residual features [14]. Our methodology results in a comparable cross entropy loss, and qualitatively similar leak probability predictions while not needing the explicit training step of SOTA alternatives. Future work on the methodology could on the one hand focus on alternative strategies to regularize the GNB classifier, and on the other hand predicting the occurrence of multiple leaks in the WDN.

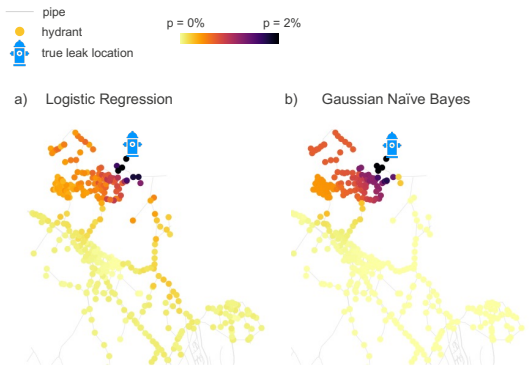

Fig. 6. Leak probabilities for leak experiment 1. Using our GNB classification methodology (b), compared with the results of a Elastic-Net logistic regression model (a) which was trained on generated pressure head residual features.

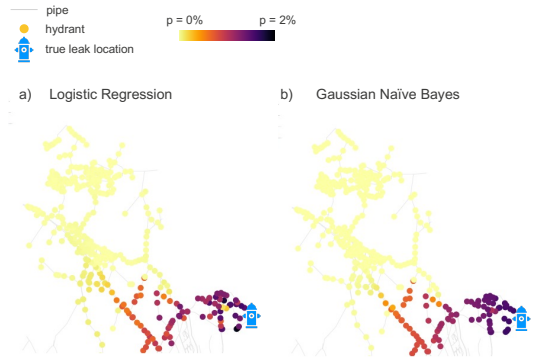

Fig. 7. Leak probabilities for leak experiment 2. Using our GNB classification methodology (b), compared with the results of a Elastic-Net logistic regression model (a) which was trained on generated pressure head residual features.

Acknowledgment

We thank P. J. Haest and J. Debaenst for their research assistance, and De Watergroep for sharing data. We thank HydroScan for setting up LeakRedux™ for the leak detection phase to determine the start and magnitude of leaks.

References

[1] M. Javadiha, J. Blesa, A. Soldevila, and V. Puig, "Leak localization in water distribution networks using deep learning," in *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, S. Elloumi, Ed., 2019, pp. 1426–1431.

[2] R. Puust, Z. Kapelan, D. Savic, and T. Koppel, "A review of methods for leakage management in pipe networks," *Urban Water J.*, vol. 7, no. 1, pp. 25–45, 2010.

[3] J. Kang, Y.-J. Park, J. Lee, S.-H. Wang, and D.-S. Eom, "Novel leakage detection by ensemble cnn-svm and graph-based localization in water distribution systems," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4279–4289, 2018.

[4] B. Arifin, Z. Li, S. L. Shah, G. A. Meyer, and A. Colin, "A novel data-driven leak detection and localization algorithm using the kantorovich distance," *Comput. Chem. Eng.*, vol. 108, pp. 300–313, 2018.

[5] E. G. Mohammed, E. B. Zeleke, and S. L. Abebe, "Water leakage detection and localization using hydraulic modeling and classification," *J. Hydroinform.*, vol. 23, no. 4, pp. 782–794, 2021.

[6] M. Quiñones-Grueiro, C. Verde, A. Prieto-Moreno, and O. Llanes-Santiago, "An unsupervised approach to leak detection and location in water distribution networks," *Int. J. Appl. Math.*, vol. 28, no. 2, pp. 283–295, 2018.

[7] J. Blesa and R. Pérez, "Modelling uncertainty for leak localization in water networks," *IFAC-Pap.*, vol. 51, no. 24, pp. 730–735, 2018.

[8] R. Perez, G. Sanz, V. Puig, J. Quevedo, M. A. Cuguero Escofet, F. Nejjari, J. Meseguer, G. Cembrano, J. M. Mirats Tur, and R. Sarrate, "Leak localization in water networks: A model-based methodology using pressure sensors applied to a real network in barcelona," *IEEE Control Syst.*, vol. 34, no. 4, pp. 24–36, 2014.

[9] M. Zhou, Y. Yang, Y. Xu, Y. Hu, Y. Cai, J. Lin, and H. Pan, "A pipeline leak detection and localization approach based on ensemble tl1dcnn," *IEEE Access*, vol. 9, pp. 47 565–47 578, 2021.

[10] A. Navarro, O. Begovich, J. Delgado-Aguiñaga, and J. Sánchez, "Real time leak isolation in pipelines based on a time delay neural network," in *2019 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, J. Cerda-Jacobo, Ed., 2019, pp. 1–6.

[11] D. Zaman, M. K. Tiwari, A. K. Gupta, and D. Sen, "A review of leakage detection strategies for pressurised pipeline in steady-state," *Eng. Fail. Anal.*, vol. 109, p. 104264, 2020.

[12] A. Soldevila, J. Blesa, S. Tornil-Sin, E. Duviella, R. M. Fernandez-Canti, and V. Puig, "Leak localization in water distribution networks using a mixed model-based/data-driven approach," *Control Eng. Pract.*, vol. 55, pp. 162–173, 2016.

[13] X. Hu, Y. Han, B. Yu, Z. Geng, and J. Fan, "Novel leakage detection and water loss management of urban water supply network using multiscale neural networks," *J. Clean. Prod.*, vol. 278, p. 123611, 2021.

[14] G. Mazaev, M. Weyns, F. Vancoillie, G. Vaes, F. Ongenae, and S. V. Hoecke, "Probabilistic leak localization in water distribution networks using a hybrid data-driven and model-based approach," *Water Supply*, 2022, in submission.

[15] K. A. Klise, M. Bynum, D. Moriarty, and R. Murray, "A software framework for assessing the resilience of drinking water systems to disasters with an example earthquake case study," *Environ. Model. Softw.*, vol. 95, pp. 420–431, 2017.

[16] Z. Fereidooni, H. Tahayori, and A. Bahadori-Jahromi, "A hybrid model-based method for leak detection in large scale water distribution networks," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, pp. 1613–1629, 2021.

[17] M. Quiñones-Grueiro, J. M. Bernal-de Lázaro, C. Verde, A. Prieto-Moreno, and O. Llanes-Santiago, "Comparison of classifiers for leak location in water distribution networks," *IFAC-PapersOnLine*, vol. 51, no. 24, pp. 407–413, 2018.

[18] I. Santos-Ruiz, J. Blesa, V. Puig, and F. López-Estrada, "Leak localization in water distribution networks using classifiers with cosenoidal features," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 16 697–16 702, 2020.

[19] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.