IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 3, 2021, accepted November 16, 2021, date of publication November 22, 2021,
date of current version November 29, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3129703*

# Detailed Leak Localization in Water Distribution Networks Using Random Forest Classifier and Pipe Segmentation

**IVANA LUČIN[1,2], ZORAN ČARIJA[1,2], SINIŠA DRUŽETA[1,2], AND BOŽE LUČIN[1,3]**
[1]Department of Fluid Mechanics and Computational Engineering, Faculty of Engineering, Faculty of Engineering, University of Rijeka, 51000 Rijeka, Croatia
[2]Center for Advanced Computing and Modelling, University of Rijeka, 51000 Rijeka, Croatia
[3]Flowtech d.o.o., 51000 Rijeka, Croatia

Corresponding author: Ivana Lučin (ilucin@riteh.hr)

This work was supported by the Center of Advanced Computing and Modelling, University of Rijeka for Providing Computing Resources.

**ABSTRACT** In this paper, a Random Forest classifier was used to predict leak locations for two differently sized water distribution networks based on pressure sensor measurements. The prediction model is trained on simulated leak scenarios with randomly chosen parameters - leak location, leak size, and base node demand uncertainty. Leak localization methods found in literature that rely on numerical simulations can only predict network nodes as leak nodes; however, since a leak can occur at any point along a pipe segment, additional spatial discretization of suspect pipe is proposed in this paper. It was observed that pipe segmentation of the whole network is a non-feasible approach since it rapidly increases the number of potential leak locations, consequently increasing the complexity of the prediction model. Therefore, a novel approach is proposed, in which a prediction model is trained on scenarios with leaks occurring in original network nodes only, but with its accuracy assessed against pressure sensor measurements from scenarios in which leaks occur in points between network nodes. It was observed that this approach can successfully narrow down the suspect leak area and, followed by additional segmentation of that network area and subsequent prediction, a precise leak localization can be achieved. The proposed approach enables incorporation of various uncertainties by simulating leak scenarios under different conditions. Investigation of leak size uncertainty and base demand variation showed that several different scenarios can produce similar sensor measurements which makes it difficult to unambiguously determine leak location using the prediction model. Therefore, future approaches of coupling prediction modeling with optimization methods are proposed.

**INDEX TERMS** Leak localization, pipe segmentation, prediction modeling, random forest, water distribution networks.

## I. INTRODUCTION

Leaks in water distribution networks can cause considerable losses, especially in older water distribution networks where considerable investments are needed for restoration. Smaller leaks can remain undetected for longer periods causing considerable water losses over time. Also, in the case of older water distribution networks rapid progression of leak size can eventually cause pipe burst which leads to water outages for end users. Therefore, a number of different techniques are being used to detect and localize leaks. These methods can be divided into hardware-based and software-based methods. Hardware-based methods use in situ visual observations or measurements. Software-based methods rely on different software for leak detection analysis. Since some methods have been developed for specialized applications depending on the transporting fluid (water, oil, gas, etc.) and location of the pipeline (water distribution network, facility, housing, etc.), a number of papers analyzed the advantages and limitations of the proposed methods and an overview of some of these methods is given in papers [1]–[5].

Software-based methods can be further divided into transient-based methods, model-based methods, and data-driven methods. Transient-based methods rely on analysis of transient pressure wave that occurs when leakage happens. For model-based methods, estimated pressure values are obtained from simulation with no leaks and in-field measured pressure values are compared, i.e. subtracted from

The associate editor coordinating the review of this manuscript and approving it for publication was Ahmed Farouk.

estimated pressure values. Obtained residuals are evaluated and if residuals are above the chosen threshold it is considered that a leak is present. Data-driven methods rely on statistical analysis and processing of raw sensor measurement data to obtain information about the presence of leaks and possible locations.

The main problem with the model-based approach is the assumption of the model being a good representation of the network. Water distribution networks have a lot of uncertainties that need to be taken into consideration, such as demand uncertainties, sensor measurement imperfections, pipe diameter uncertainties, etc. Thus, the model-based approach cannot capture all these parameters. The data-driven approach using raw sensor measurements could incorporate all these variations, but the main problem is the number of leak events which are rather sparse. Since the amount of data is small compared to the amount of data needed for efficiently employing machine learning algorithms, models can be advanced by incorporating uncertainties through simulations with varying parameters which can produce additional data.

Machine learning has been used for a variety of water distribution system applications. Prediction of failure of water mains was investigated in [6] where artificial neural network (ANN), ridge regression, and ensemble decision tree were used. Different machine learning algorithms have been explored for the prediction of leak locations in pipelines, such as convolutional neural network (CNN) [7], [8] and ANN [9], [10]. In [11] support vector machine (SVM) method was used to predict leaks in wall-mounted pipelines.

When considering water distribution networks, in [12], a deep learning model based on additional pressure meters installed on optimal places was used to identify pipe burst locations. In [13], SVM was used for prediction of leak size and location based on pressure sensors gathered from EPANET simulations for small size leakages. In [14], leakage detection was conducted for 1500 m × 1500 m experimental network using principal component analysis (PCA) and SVM. In [15], model-based method was used to identify leak event and data-driven approach using k-Nearest Neighbors (k-NN) classifier was used in the second stage to determine leak location. In the further study [16] Bayesian classifier was used with improved localization accuracy. Both methods were applied to real water distribution network case studies. In [17], unsupervised principal component analysis (PCA) approach for leak detection was conducted for the Hanoi distribution network. In [18], Kriging method was used to estimate pressure measurements in the whole network based on the limited number of sensor measurements and classification methods were used to determine leak location. It was shown that the accuracy of the proposed method was very low for some sensor layouts due to Kriging interpolation error. In [19], detection and localization of multiple leak locations were explored. SVM was used as a classifier for leak detection using the residual method and a statistical method was used for leak localization in the Hanoi network. All mentioned papers assume possible leak locations only in network nodes.

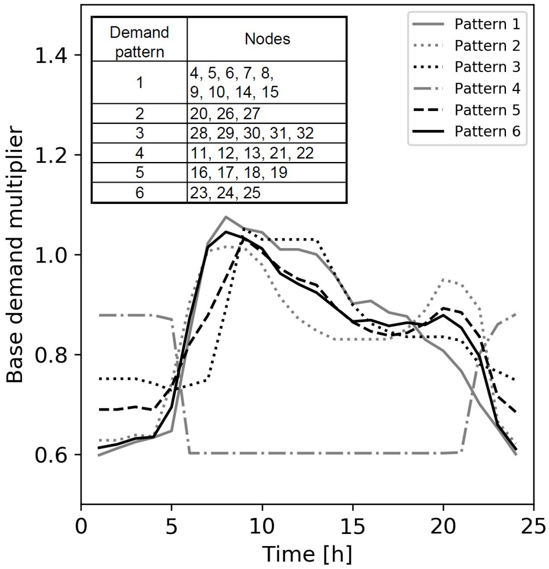
In order to increase the number of input data, in previous work [20] it was proposed that a great number of leak scenarios can be generated by simulating different leak locations and leak sizes under different demand uncertanties. The machine learning approach for leak localization was investigated for variously sized water distribution networks, various demand ranges, and various sensor placements. However, considerable simplification was made insomuch that the prediction model was trained with simulated scenarios in which leak locations occur only in network nodes while in reality leaks can occur at any point along a pipe segment. Thus, in this paper, an approach with pipe segmentation in suspect areas is investigated. The idea is taken from the adaptive mesh refinement approach used in computational fluid dynamics (CFD) simulations, where the area of interest is refined with additional numerical nodes in order to increase the accuracy of results. An alternative approach of fault zone identification has been used in work by [21] and [22]. However, that approach could be problematic for leak locations at the borders of leak zones since water distribution network needs to be divided into zones before using leak localization method. The approach proposed in this paper identifies suspect nodes from machine learning prediction model, which then serve as indicators for pipes that need to be further explored using segmentation. Therefore a possible leak area is adjusted for each leak event based on prediction results.

In the first part of this paper, it is investigated whether a prediction model trained only on simulations with leak locations in network nodes can successfully predict leaks that occur in-between network nodes. Two differently-sized water distribution networks, Hanoi and Net3 were used for this, coupled with various sensor layouts, leak sizes, and demands. Furthermore, the accuracy of sequential prediction models in predicting leak location was investigated. The prediction model performance is investigated when several most-suspect nodes are considered and segmentation of pipes near those suspect nodes is performed. The subsequent prediction model is trained on scenarios with leak locations in most-suspect network nodes and in nodes added through pipe segmentation from the previous stage. Limitations of the proposed method and future work are presented in the discussion section.

## II. METHODOLOGY

### A. PROBLEM STATEMENT

Leak localization methods based on machine learning methods require considerable amount of data for model training. Since the measurements for real leak events are rather sparse, additional data can be obtained by simulating different leak scenarios. For this purpose, leak scenarios were simulated using EPANET version 2.0.12. [23] with various leak scenario parameters. Leak location, leak size, and node demands were chosen randomly to cover a wide range of possible leak events. Typically it is assumed that water distribution network models are calibrated and that leaks can occur only



| Demand pattern | Nodes |
|---|---|
| 1 | 4, 5, 6, 7, 8, 9, 10, 14, 15 |
| 2 | 20, 26, 27 |
| 3 | 28, 29, 30, 31, 32 |
| 4 | 11, 12, 13, 21, 22 |
| 5 | 16, 17, 18, 19 |
| 6 | 23, 24, 25 |

**FIGURE 1. Demand patterns used for Hanoi network.**

in network nodes. The latter assumption can be problematic for water distribution networks with longer pipe segments since localization will be a very rough estimate. Therefore, additional pipe segmentation is introduced which divides a pipe into smaller sections, allowing better leak localization. Random Forest machine learning algorithm is trained with pressure sensor measurements from simulated scenarios and is then employed to determine most suspect leak locations.

### B. WATER DISTRIBUTION NETWORKS

The investigated water distribution networks are small-sized Hanoi network and medium-sized Net3 network. Hanoi (Vietnam) network with 31 nodes was obtained from The Centre for Water Systems (CWS) at the University of Exeter [24]. For Hanoi network, demand patterns as described in [17] are adopted (Figure 1). Net3 network is an EPANET example network for dual-source system that changes over time, consisting of 92 nodes. For both networks simulation time was 24 h, hydraulic time step was 10 min and report time step 1 h. To generate a wide range of possible leak scenarios, emitter coefficient and leak location were chosen randomly. Additionally, to incorporate demand variation, it was randomly decided whether node base demand was to be changed or not. If it was chosen to be changed, base demand was increased or decreased by randomly chosen percentage in the range ±2.5% or ±5%.

For each water distribution network, two different sensor layouts were considered. For Hanoi network, the first layout has two sensors located in network nodes 14 and 30, as given in [15], and the second layout has three sensors located in network nodes 8, 20, and 31, as given in [17] (Figure 2). For Net3 network, the first layout has four sensors located in network nodes 117, 143, 181, and 213, and the second layout has two sensors located in network nodes 117 and 181 (Figure 3).

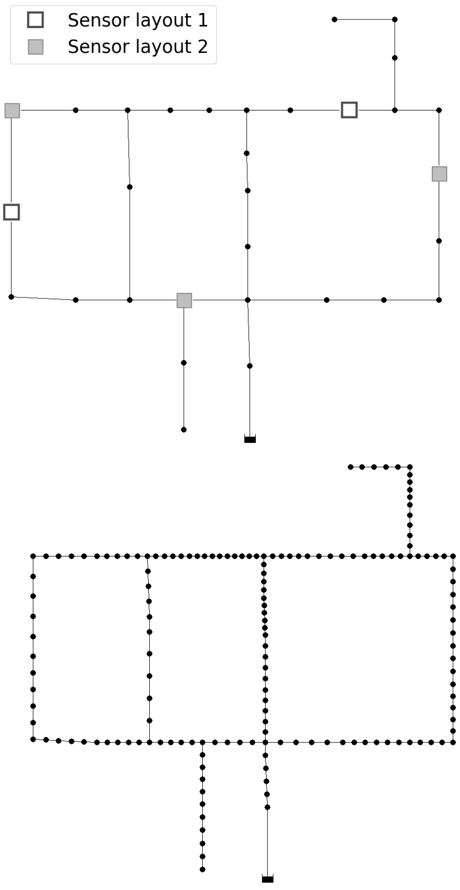


**FIGURE 2. Hanoi network, original with indicated sensor locations (above), and after pipe segmentation with 5 segments per pipe (below).**

### C. PIPE SEGMENTATION

Discretization of water distribution network pipes was achieved by inserting additional network nodes, where each pipe was split on 5 segments of equal length, resulting in additional 4 nodes for each pipe (Figure 2). Although it would be more beneficial to define a fixed segment length, a fixed number of segments was used as a methodological simplification.

To investigate machine learning efficiency in the localization of leak locations in pipe segments, three different models were analyzed. Model 1 was trained and tested on leak scenarios with leak locations in original network nodes. Model 2 was trained and tested on leak scenarios with leaks located both in network nodes and refinement nodes, resulting in a significantly increased number of ML output classes. Finally, Model 3 was trained on scenarios with leaks in original network nodes, but it was then tested for scenarios in which leak locations can be both in network nodes and refinement nodes.

Flowcharts of the proposed models can be observed in Figure 4. Depending on considered model leak node $N_i$ is chosen from original network nodes $N_i \in \{N_0^o, \ldots, N_{no}^o\}$ where superscript $o$ denotes original network nodes, or from original network nodes and additional nodes generated from

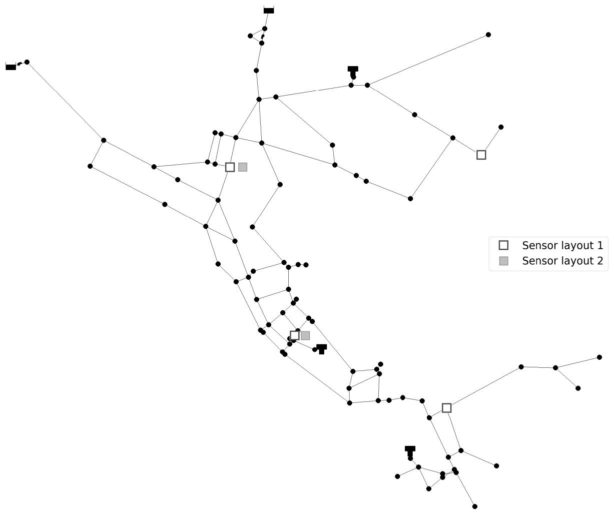


**FIGURE 3. Net3 network with indicated sensor locations for considered sensor layouts.**

segmentation $N_i \in \{N_0^o, \ldots, N_{no}^o, N_0^s, \ldots, N_{ns}^s\}$ where superscript $s$ denotes segmentation nodes, subscript $no$ denotes total number of original network nodes and $ns$ total number of segmentation nodes. The sensor measurements $S_i \in \{S_0(t), \ldots, S_n(t)\}$, were $n$ indicates total number of sensors for considered sensor layout, were recorded through time $t$, namely 25 timesteps in all considered cases. Since the model 3 is trained only on the original network nodes it cannot possibly predict a refinement node. Thus the refinement nodes are considered to be predicted correctly if their nearest original network node $N_i \in \{N_0^o, \ldots, N_{no}^o\}$ was predicted. This simulates a most realistic scenario where leaks can occur anywhere in the pipe segment, however, the model can be trained only with scenarios with leaks in network nodes we have in the model.

### D. RANDOM FOREST CLASSIFIER

Machine learning (ML) algorithms are being used to find underlying correlations or patterns from obtained data. This ability enables machine learning algorithms to provide a prediction for unseen data, which can be categorized into regression and classification problems. Regression algorithms are designed to provide a prediction of the exact value of the output variable, while classification algorithms separate data into logical groups, i.e. classes.

Random Forest classifier was first proposed by [25] and is an ensemble type of algorithm based on multiple decision trees which are created as independent prediction models. Decision trees (DT) are constructed in a form of flowchart structure, where nodes represent attributes used for outcome prediction. Based on feature values a decision is made at each node and ultimately based on these decisions classification is reached. Each tree is defined with tree depth parameter which defines how many splits can be made before making a prediction. Random Forest uses bootstrap and aggregation methods to obtain unique data subsets for the training of each decision tree and to ultimately count the class with the most predictions. Increased number of trees increases the precision of the classifier, albeit also increasing its complexity. The problem considered in this paper is the classification problem since each potential leak node represents one class, thus Random Forest classifier was adopted as a suitable ML method. Random Forest classifier implementation in the Python library Scikit-learn [26] version 0.20.3 was used.

The dataset is composed of 500 000 inputs, with training-testing split 70%-30%, resulting in 350 000 training records and 150 000 testing records. It was observed in [20] that a smaller timestep only slightly increases prediction accuracy so timestep of 1 hr was adopted in order to reduce number of features and reduce computational time.

Grid search optimization of Random Forest parameters was conducted for Hanoi network with 100 000 inputs with leak coefficient range $10 \ldots 15$ and with $\pm 2.5\%$ demand variation in order to find optimal number of estimators (trees), maximum depth, and minimum number of samples required to split an internal node. It was found that the optimal minimum number of samples required to split an internal node is 2, the optimal maximum depth of the tree is 20, and the optimal number of estimators (i.e. trees) is 200. These parameters are kept constant for all investigated prediction models. Other Random Forest parameters were kept at default values of the Scikit-learn implementation. For each prediction model, five runs were conducted to consider the influence of prediction model parameter randomness and average accuracy values are reported. Additionally, model accuracy was measured for true leak node presence in top 3 and top 5 suspect network nodes with greatest prediction certainties. Even if true leak node is not correctly predicted, presence of true leak node in top 3 or top 5 most suspect nodes considerably narrows down the area of leak location.

## III. RESULTS

### A. EFFECT OF PIPE SEGMENTATION

Hanoi network with two sensors, emitter coefficient range $10 \ldots 15$ and different demand variations was investigated first. In Model 1, where leaks can occur only in the original network nodes, 31 prediction classes were obtained. For Model 2 each pipe segment is divided into 5 segments of equal length, resulting in 163 prediction classes. Although Model 3 was used for predicting leak scenarios on segmented network of Model 2, it was trained on leak scenarios used for Model 1. Thus, in Model 3 the 31 prediction classes corresponding to the original network nodes were used, with leaks in the 135 segmentation nodes expected to be classified as leaks in their nearest original network nodes.

Results for the conducted investigation are presented in Table 1. It can be observed that with the increase of demand variation, model accuracy considerably decreases; indicating a rapid increase of possible scenarios which are consequently difficult to predict. However, when top 3 and top 5 suspect network nodes with the greatest certainties are considered, model accuracy is high. For Model 2, where 163 network nodes are possible prediction classes, model accuracy is very

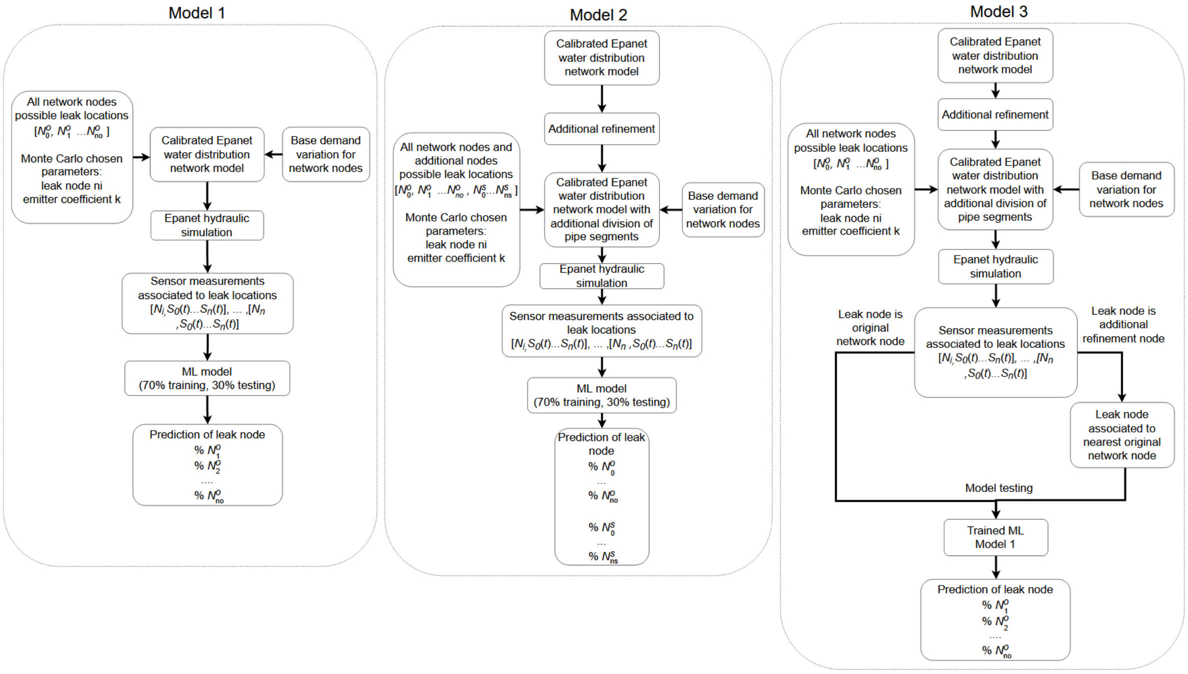

Model 1
All network nodes possible leak locations $[N_0^o, N_1^o \dots N_{no}^o]$
Monte Carlo chosen parameters: leak node ni emitter coefficient k
Calibrated Epanet water distribution network model
Base demand variation for network nodes
Epanet hydraulic simulation
Sensor measurements associated to leak locations $[N_i S_0(t) \dots S_n(t)], \dots, [N_n, S_0(t) \dots S_n(t)]$
ML model (70% training, 30% testing)
Prediction of leak node % $N_1^o$ % $N_2^o$ .... % $N_{no}^o$
Model 2
Calibrated Epanet water distribution network model
Additional refinement
All network nodes and additional nodes possible leak locations $[N_0^o, N_1^o \dots N_{no}^o, N_0^s \dots N_{ns}^s]$
Monte Carlo chosen parameters: leak node ni emitter coefficient k
Calibrated Epanet water distribution network model with additional division of pipe segments
Base demand variation for network nodes
Epanet hydraulic simulation
Sensor measurements associated to leak locations $[N_i S_0(t) \dots S_n(t)], \dots, [N_n, S_0(t) \dots S_n(t)]$
ML model (70% training, 30% testing)
Prediction of leak node % $N_0^o$ ... % $N_{no}^o$ % $N_0^s$ ... % $N_{ns}^s$
Model 3
Calibrated Epanet water distribution network model
Additional refinement
All network nodes possible leak locations $[N_0^o, N_1^o \dots N_{no}^o]$
Monte Carlo chosen parameters: leak node ni emitter coefficient k
Calibrated Epanet water distribution network model with additional division of pipe segments
Base demand variation for network nodes
Epanet hydraulic simulation
Leak node is original network node
Sensor measurements associated to leak locations $[N_i S_0(t) \dots S_n(t)], \dots, [N_n, S_0(t) \dots S_n(t)]$
Leak node is additional refinement node
Leak node associated to nearest original network node
Model testing
Trained ML Model 1
Prediction of leak node % $N_1^o$ % $N_2^o$ ... % $N_{no}^o$

**FIGURE 4.** Flowcharts of the considered machine learning approaches.

low, indicating that for greater networks this approach would require even more data and computational resources, which is currently not feasible. Model 3 accuracy is reduced compared to the Model 1 approach, which is expected as segmentation nodes increase the total number of possible leak locations. Furthermore, leaking in the segmentation nodes in the middle of the pipe could provide flow patterns that could be equally similar to flow patterns produced by leaking on one or the other edge node of that pipe, thus also contributing to reduced accuracy. However, when top 3 and top 5 suspect nodes are considered, the difference in prediction accuracy for Model 1 and Model 3 shrinks to only around a couple of percents. Although the proposed ML approach demonstrates modest accuracy in predicting the exact leak locations, the proposed approach can be successfully used to narrow down the leak location.

The same investigation was conducted for Net3 network with 4 sensors, emitter coefficient range 10 . . . 15, and for different demand variation ranges. Model 1 and Model 3 are created with 92 classes, while Model 2 was also created with 5 additional segments per pipe, resulting in 544 classes altogether.

Results for Net3 are reported in Table 2. It can be observed that prediction model accuracy for the Net3 network is significantly lower than for the Hanoi network. For a model with no demand variation, it is around 7% lower than for the Hanoi network and with an increase in demand variation this decline is over 20%. This is expected, since the Net3 network has a greater number of network nodes and consequently a greater number of possible leak locations. Model 2 accuracy is very small, especially for the strongest variation of demand, as it was observed for the Hanoi network, confirming this approach is not feasible. However, although Model 3 has reduced accuracy when compared with Model 1, when considering top 3 and top 5 nodes the accuracy of Model 3 comes very close to the accuracy of Model 1, indicating that the Model 3 approach could be successfully used in a real leak scenario.

**TABLE 1.** Influence of Hanoi network refinement on prediction model accuracy for emitter coefficient range 10 . . . 15 for different ranges of demand variation. Results are average of 5 runs with 500 000 inputs (350 000 training records).

| | Demand variation | | |
|---|---|---|---|
| | None | ±2.5% | ±5% |
| | Model 1 | | |
| Accuracy | 100% | 82% | 69% |
| Top 3 | 100% | 96% | 91% |
| Top 5 | 100% | 99% | 97% |
| | Model 2 | | |
| Accuracy | 99% | 36% | 21% |
| Top 3 | 99% | 68% | 49% |
| Top 5 | 99% | 80% | 64% |
| | Model 3 | | |
| Accuracy | 82% | 68% | 57% |
| Top 3 | 97% | 94% | 88% |
| Top 5 | 99% | 98% | 96% |

Considering these results, only Model 1 and Model 3 will be considered in further research.

## B. SENSOR AND EMITTER COEFFICIENT INFLUENCE

The investigation was conducted for various sensor placements, number of sensors and emitter coefficient ranges. The results for the Hanoi network are presented in Table 3. It can be observed that overall prediction model accuracy decreases with greater coefficient range. This is expected since a greater coefficient range increases the size of the problem solution space. On the other hand, with a greater number of sensors,

**TABLE 2.** Influence of Net3 network refinement on prediction model accuracy for emitter coefficient range 10 . . . 15 for different ranges of demand variation. Results are average of 5 runs with 500 000 inputs (350 000 training records).

| | Demand variation | | |
|---|---|---|---|
| | None | ±2.5% | ±5% |
| | Model 1 | | |
| Accuracy | 93% | 58% | 46% |
| Top 3 | 99% | 89% | 77% |
| Top 5 | 100% | 97% | 88% |
| | Model 2 | | |
| Accuracy | 77% | 22% | 13% |
| Top 3 | 87% | 42% | 28% |
| Top 5 | 92% | 54% | 38% |
| | Model 3 | | |
| Accuracy | 69% | 45% | 36% |
| Top 3 | 93% | 83% | 70% |
| Top 5 | 98% | 93% | 84% |

**TABLE 3.** Prediction model accuracy for Hanoi network for various emitter coefficient ranges, sensor layouts and demand variations for model 3.

| | | Demand variation | | |
|---|---|---|---|---|
| | | None | ±2.5% | ±5% |
| Emitter coeff. | 2 sensors | | | |
| 10..15 | Model 1 | 100% | 80% | 67% |
| | Model 3 | 81% | 66% | 56% |
| 5..15 | Model 1 | 100% | 68% | 53% |
| | Model 3 | 85% | 57% | 45% |
| Emitter coeff. | 3 sensors | | | |
| 10..15 | Model 1 | 100% | 82% | 69% |
| | Model 3 | 82% | 68% | 57% |
| 5..15 | Model 1 | 100% | 73% | 57% |
| | Model 3 | 86% | 60% | 48% |

prediction model accuracy slightly increases. Additionally, the greatest difference in Model 1 and Model 3 accuracy appears for scenarios with no demand variation, ranging from 15% to 19%. However, as demand variation increases, the accuracy difference falls to 8 . . . 12%.

The results for Net3 network are presented in Table 4. Same as in the Hanoi network case, with a greater range of emitter coefficient both Model 1 and Model 3 accuracy decrease, for both sensor layouts. Same as in the Hanoi case, as demand variation increases the difference between Model 1 and Model 3 accuracy decreases and again the greatest difference in model accuracy is for no demand variation.

### C. PIPE SEGMENT SEGMENTATION INFLUENCE

In order to investigate pipe segmentation influence in the Model 3 approach, three different discretizations are considered for the Net3 network with 4 sensors. Pipes were divided into 3, 5, and 11 segments, resulting in 318, 544, and 1222 possible leak locations, respectively. The results are presented in Table 5. It can be observed that a finer network segmentation slightly reduces model accuracy, which is entirely expected since the number of prediction classes rises with greater refinement. Also, it is expected that at some point further refinement would lead to scenarios with different leak nodes but almost identical pressure readings, since these nodes may happen to be situated very close to each other. However, the rather small decline in accuracy indicates that the proposed approach can be successfully used to narrow down a leak location.

**TABLE 4.** Prediction model accuracy for Net3 network for various emitter coefficient ranges, sensor layouts and demand variations.

| | | Demand variation | | |
|---|---|---|---|---|
| | | None | ±2.5% | ±5% |
| Emitter coeff. | 2 sensors | | | |
| 10..15 | Model 1 | 90% | 50% | 36% |
| | Model 3 | 64% | 41% | 30% |
| 5..15 | Model 1 | 85% | 41% | 28% |
| | Model 3 | 63% | 34% | 23% |
| Emitter coeff. | 4 sensors | | | |
| 10..15 | Model 1 | 93% | 58% | 46% |
| | Model 3 | 69% | 48% | 37% |
| 5..15 | Model 1 | 89% | 49% | 36% |
| | Model 3 | 67% | 41% | 30% |

**TABLE 5.** Prediction model 3 accuracy for Net3 network for various number of pipe segments and emitter coefficient ranges.

| | Segments | Demand variation | | |
|---|---|---|---|---|
| Emitter coeff. | per pipe | None | ±2.5% | ±5% |
| 10..15 | 3 | 70% | 48% | 38% |
| | 5 | 69% | 48% | 37% |
| | 11 | 67% | 48% | 37% |
| 5..15 | 3 | 68% | 41% | 30% |
| | 5 | 67% | 41% | 30% |
| | 11 | 66% | 41% | 30% |

### D. ACCURACY IMPROVEMENT

The number of top suspect nodes which need to be considered to achieve 99% model accuracy was investigated to increase accuracy of the prediction model. This approach was already used in [27] to localize the source of pollution and similarly in [13] where the correlation between accuracy and distance between predicted and actual leak node was presented. In this way, a considerable number of network nodes is eliminated, thus the leak area can be localized with considerable certainty even for sparse sensor placement and greatest demand variation.

Number of needed top nodes for Hanoi network is presented in Table 6. It can be observed for Model 1 that with the increase in demand variation, a greater number of top nodes needs to be considered to achieve 99% accuracy; however, considerable localization is achieved even for the strongest demand variation. Similar behavior can be observed with Model 3, where the greatest number of top nodes needs to be considered for the greatest demand variation. Also, a number of top nodes comparing to Model 1 is slightly greater, which is expected. In Figure 5, the increase of model accuracy with the increase of considered top nodes is illustrated. It can be observed that for all models the accuracy of 90% is surpassed when using only top 4 nodes. Additionally, a rapid increase in prediction model accuracy is observed when including the first several top nodes. However, after some threshold the additional nodes in the top list only slightly improve the overall model accuracy.

This kind of investigation has also been conducted for the Net3 network, and the results are presented in Table 7. The number of top nodes is greatest for Model 3 and for stronger demand variation, which is expected and consistent with Hanoi results. It must be noted that even for the worst performing model, with emitter coefficient range 5 . . . 15 and demand variation ±5%, 32 top nodes represent only 35% of

**TABLE 6. Number of top nodes needed to achieve 99% accuracy for Hanoi network with various emitter coefficient ranges, sensor layouts and demand variations.**

| | | Demand variation | |
|---|---|---|---|
| | | ±2.5% | ±5% |
| Emitter coeff. | 2 sensors | | |
| 10..15 | Model 1 | 5 | 6 |
| | Model 3 | 5 | 7 |
| 5..15 | Model 1 | 7 | 9 |
| | Model 3 | 8 | 10 |
| Emitter coeff. | 3 sensors | | |
| 10..15 | Model 1 | 5 | 7 |
| | Model 3 | 6 | 8 |
| 5..15 | Model 1 | 7 | 9 |
| | Model 3 | 8 | 10 |

**TABLE 7. Number of top nodes needed to achieve 99% accuracy for Net3 network for various emitter coefficient ranges and demand variations.**

| | | Demand variation | |
|---|---|---|---|
| | | ±2.5% | ±5% |
| Emitter coeff. | 4 sensors | | |
| 10..15 | Model 1 | 8 | 21 |
| | Model 3 | 12 | 24 |
| 5..15 | Model 1 | 15 | 29 |
| | Model 3 | 19 | 32 |

all network nodes, which is still a considerable localization. Additionally, it must be taken into consideration that the chosen 99% accuracy threshold is very high, where the strong model accuracy manifests even for the smaller number of top nodes (Figure 6). To further evaluate the proposed model, the sequential prediction modeling approach is evaluated in the next section.

### *E. REALISTIC SCENARIO TESTING*

To further evaluate the proposed ML approach, an investigation was conducted for a simulated case on Net3 network with 30 records which represent 30 different days. Scenarios are generated with fixed leak location and leak coefficient, but with different demands in network nodes obtained through base demand variation of ±2.5%. Two different leak locations were chosen, first with leak location in network node 159 (Figure 7) with emitter coefficient set to 10, and second with leak location in a pipe segment between nodes 205 and 207 (Figure 8) and with emitter coefficient set to 15. The initial prediction was made using Model 1 with emitter coefficient range 10 . . . 15 and base demand variation of ±2.5%. From previous investigation (Table 7) it was observed that when leak locations in pipe segment nodes are allowed, the top 12 nodes achieve 99% accuracy, thus 12 nodes with the greatest prediction model certainty are considered for further segmentation and secondary Model 3 predictions.

For each of the 30 records different certainties are obtained, i.e. the top 12 nodes could be different for each record. Therefore, the average value of all 30 certainties for each node was chosen as a measure for choosing the top 12 nodes with the greatest certainty. For leak node 159, the greatest model certainty is obtained for true leak location, where for leak node in pipe segment between nodes 205 and 207 the greatest certainty is obtained for leak location 207 which is the edge node of the considered pipe segment. Suspect nodes for both considered cases are presented in Figures 7 and 8, with indicated top 3 nodes with greatest certainty. It can be observed that the top 3 nodes always include true leak location, together with network nodes in the immediate vicinity of the true leak location.

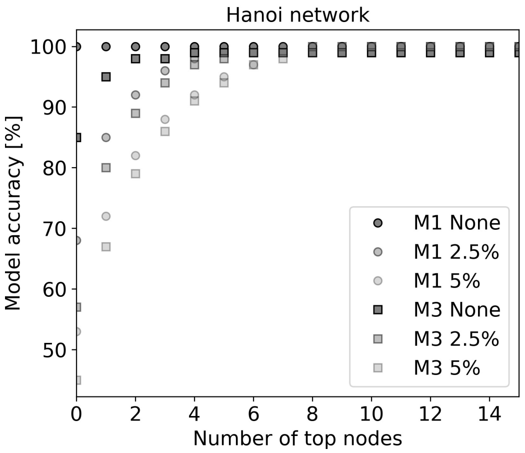


**FIGURE 5. Influence of the number of top nodes on prediction model accuracy for Hanoi network with two sensors and emitter coefficient in range 5 . . . 15.**

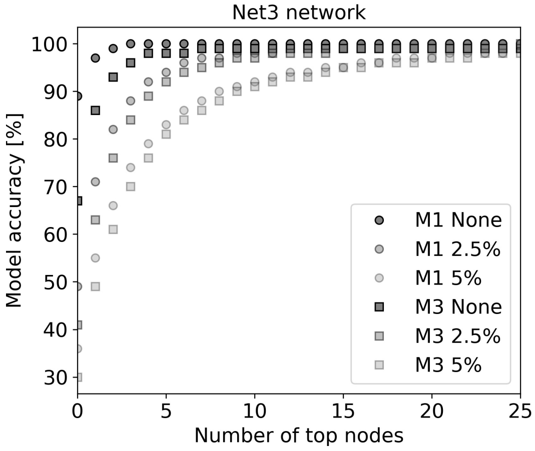


**FIGURE 6. Influence of the number of top nodes on prediction model accuracy for Net3 network with 4 sensors layout and emitter coefficient range 5 . . . 15.**

For the next stage, additional pipe segmentation was performed around these top 12 nodes and a prediction model was created where possible leak locations were the top 12 network nodes plus the newly inserted nodes. At this stage, for leak location 159, the most suspect node was node 60, and the second candidate node was node 159 which is the true leak node. For leak location in pipe segment between nodes 205 and 207, the most suspect node was node right next to the true leak node and the second candidate was the true leak node. Top 3 most suspect nodes for both considered cases can be observed in Figures 9 and 10.

The third sequential prediction model was trained also on the top 12 nodes with the greatest average certainty from the previous stage. Both considered cases have true leak location as the second most suspect node. Additionally, from

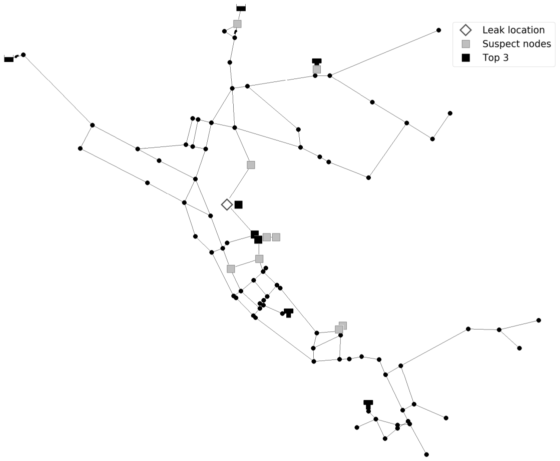


**FIGURE 7.** Net3 realistic scenario testing for source node 159 with indicated suspect nodes and top 3 nodes at first stage.

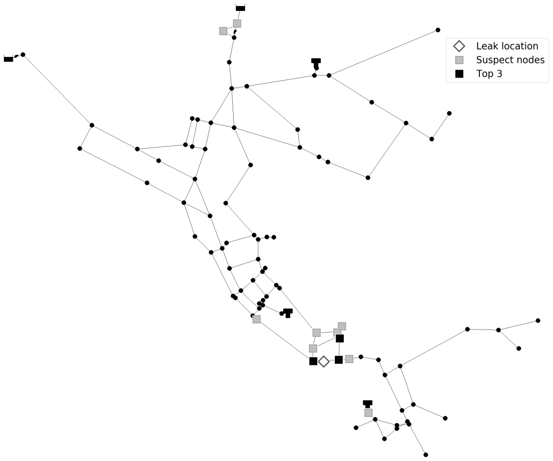


**FIGURE 8.** Net3 realistic scenario testing for source node between network nodes 205 and 207 with indicated suspect nodes and top 3 nodes at first stage.

Figures 9 and 10 variation in top 3 most suspect nodes can be observed, showing that an unambiguous solution cannot be obtained. This indicates that for different leak locations, demands, and emitter coefficients, still a very similar pressure measurement can be obtained. In other words, there are multiple solutions to the problem. It is shown that the prediction model can efficiently localize leak areas for sparse sensor placement for leak locations which can occur anywhere in pipe segments. However, due to wide range of leak scenarios that are used for prediction model training, a prediction model for fine localization may not be able to provide a single solution.

## IV. DISCUSSION

It is shown that the proposed ML approach can be successfully used for localization of leak area under demand uncertainty, for different sized networks, and for different sensor placement layouts. ML model for segmented network pipes was investigated to take into consideration that leaks can occur anywhere along a pipe, but it was shown to be an unfeasible approach. Any pipe segmentation considerably

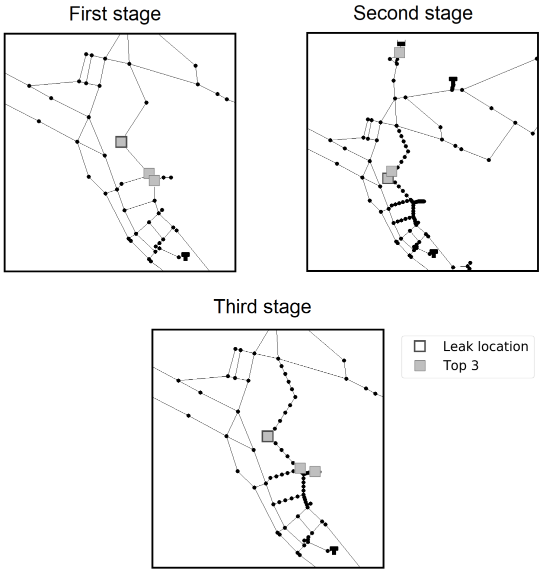


**FIGURE 9.** Realistic scenario testing for Net3 network for leak location in network node 159 with indicated top 3 nodes through refinement stages.

increases the number of network nodes, i.e. number of prediction classes, with consequently rapidly increasing computational complexity. Additionally, a greater number of inputs is required, which is a considerable problem for greater networks. However, it seems that leaks that occur in pipe segments can be successfully localized with a prediction model trained only on scenarios generated for original network nodes, especially when several top most suspect nodes are considered. It was also observed that regardless of pipe refinement, similar prediction accuracy can be obtained. However, as was mentioned before, a simplification was made where all pipes, regardless of their length, had the same number of divisions. Therefore in future work, fixed lengths for additional refinement nodes should be explored to further explore the presented approach and align the proposed technique with practical purposes.

Sequential prediction models were tested, where the first prediction model specified area for further segmentation, and subsequent models were used to find the exact leak location. It was observed that ML has a problem with detecting fine differences in leak scenarios; the true leak location was always in top nodes but was not always the node with the greatest model certainty. This can be explained by the fact that machine learning models need to cover a large span of scenarios (different demands, different leak sizes, etc.), thus it is reasonably expected that several equally good solutions exist. Similar observation was made in [15] where some leak locations were grouped in single classes, since distinction between locations could not be made.

In further research, coupling of ML and optimization methods needs to be explored. Genetic algorithm (GA) was explored in [28] for leak localization using the inverse transient method for a network with 7 nodes. The main problem with optimization methods in water distribution networks is the network node variable, which is a categorical variable

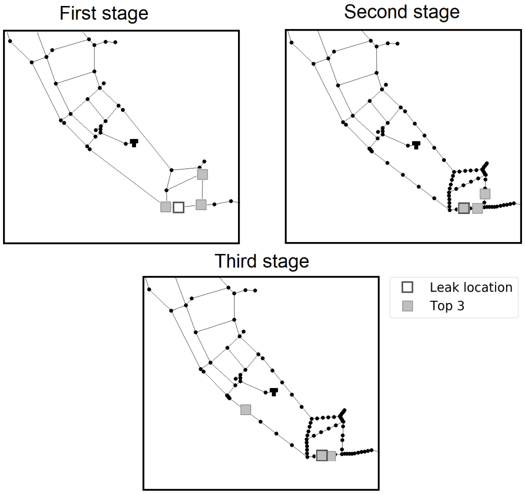


**FIGURE 10.** Realistic scenario testing for Net3 network for leak location between network nodes 205 and 207 with indicated top 3 nodes through refinement stages.

and as such makes the optimization problem very complex and computationally demanding. However, if ML is used to localize a leak area, independent optimizations for suspect nodes can be conducted and thus reduce the optimization complexity. This was successfully applied in [29] where the pollution source was localized and independent optimizations were conducted to obtain a true pollution source. Additionally, if the optimization method is to be employed, network demands could be more carefully monitored for some period, for example from 2 to 3 AM as proposed in [13], to eliminate or reduce demand variation which is shown to considerably decrease prediction accuracy.

It must be noted that Random Forest classifier was chosen due to its simplicity and since it allows for a reasonably reliable prediction without method parameter fine tuning. For example in [30] RF classifier outperformed SVM, ANN, k-NN and DT for leak detection using acoustic signals, however extensive analysis of classifier parameters was not shown. In [31] six deep neural networks structures and three RF classifier were compared for source tracking of chemical leaks and best accuracy was achieved with RF classifier. Additionally, in [32] Gradient Boosting, DT, RF, SVM and ANN models were investigated for detection of leaks in natural gas pipelines where models were tuned to ensure no false alarm. ANN and SVM showed best performance, however RF and DT were most sensitive to detect small leaks. Therefore, it can be concluded that other models such as ANN may outperform Random Forest algorithm if fine-tuning of hyper-parameters is conducted. Novel ANN methods which deal with this ANN complexity are being developed such as quantum-inspired neural network Autonomous Perceptron Model [33] which showed better performance than other algorithms, including classic ANN and RF. Therefore, extensive investigation of other machine learning algorithms should be conducted in future work to determine which classifier can provide best model accuracy for leak localization problem in water distribution networks. Dimensionality reduction methods should also be explored to reduce the number of features, consequently reducing prediction model complexity which could be important for bigger water distribution networks.

The proposed methodology could provide real-time support in water distribution network surveillance. The prediction model can be prepared with incorporated demand uncertainties, and can therefore be continuously used to detect when a single leak location is repeatedly reported. However, future work should investigate the possibility of identification of multiple leak locations, which is also most often the case. Other uncertainties should also be incorporated, such as sensor measurement uncertainties and model uncertainties such as pipe diameter and pipe roughness. Ultimately, the proposed methodology should be tested on real-world water distribution network data where all these uncertainties are present.

## V. CONCLUSION

In this paper, machine learning approach using big data obtained from computer simulations was investigated for leak localization in water distribution networks. In previous research, a simplification was made in which leaks were only occurring in network nodes and here the methodology is enhanced by allowing for leaks to occur anywhere on any network pipe. It was observed that global refinement of the network in which segmentation is performed on all pipes is not a feasible approach, since the number of potential leak locations rapidly increases and construction of a capable machine learning model is currently computationally too demanding.

However, only a small reduction in model accuracy is observed when the prediction model is trained exclusively on scenarios with leaks appearing in network nodes, while the prediction is then given for leak scenarios with leaks in pipe segments. Further investigation showed that this reduction in model accuracy can be compensated by considering several most suspect nodes. This approach significantly narrows down the leak area, especially if larger water distribution networks are considered. These observations indicate that the proposed approach could be applicable in real-world water distribution networks and further study of the proposed approach should be conducted.

In future research, additional model uncertainties regarding pipe roughness and pipe lengths should be included. Since it is observed that increasing demand uncertainty rapidly decreases model accuracy, an additional approach should also include dimensionality reduction of input data. Sequential prediction models were also explored, where further prediction models were trained using only leak scenarios for most suspect leak nodes from the previous prediction model. This approach was shown not to be beneficial since prediction models provide a generalized model, and further leak localization needs a specific solution. Coupling the proposed methodology with an optimization procedure could provide better results, which should be explored in future work.

## REFERENCES

[1] S. Datta and S. Sarkar, "A review on different pipeline fault detection methods," *J. Loss Prevention Process Ind.*, vol. 41, pp. 97–106, May 2016.

[2] A. Gupta and K. D. Kulat, "A selective literature review on leak management techniques for water distribution system," *Water Resour. Manage.*, vol. 32, no. 10, pp. 3247–3269, Aug. 2018.

[3] U. Baroudi, A. A. Al-Roubaiey, and A. Devendiran, "Pipeline leak detection systems and data fusion: A survey," *IEEE Access*, vol. 7, pp. 97426–97439, 2019.

[4] D. Zaman, M. K. Tiwari, A. K. Gupta, and D. Sen, "A review of leakage detection strategies for pressurised pipeline in steady-state," *Eng. Failure Anal.*, vol. 109, Jan. 2020, Art. no. 104264.

[5] M. I. Mohd Ismail, R. A. Dziyauddin, N. A. Ahmad Salleh, F. Muhammad-Sukki, N. Aini Bani, M. A. Mohd Izhar, and L. A. Latiff, "A review of vibration detection methods using accelerometer sensors for water pipeline leakage," *IEEE Access*, vol. 7, pp. 51965–51981, 2019.

[6] Z. Almheiri, M. Meguid, and T. Zayed, "Intelligent approaches for predicting failure of water mains," *J. Pipeline Syst. Eng. Pract.*, vol. 11, no. 4, Nov. 2020, Art. no. 04020044.

[7] M. Zhou, Z. Pan, Y. Liu, Q. Zhang, Y. Cai, and H. Pan, "Leak detection and location based on ISLMD and CNN in a pipeline," *IEEE Access*, vol. 7, pp. 30457–30464, 2019.

[8] H. Shukla and K. Piratla, "Leakage detection in water pipelines using supervised classification of acceleration signals," *Autom. Construction*, vol. 117, Sep. 2020, Art. no. 103256.

[9] J. Bohorquez, B. Alexander, A. R. Simpson, and M. F. Lambert, "Leak detection and topology identification in pipelines using fluid transients and artificial neural networks," *J. Water Resour. Planning Manage.*, vol. 146, no. 6, Jun. 2020, Art. no. 04020040.

[10] E. J. Pérez-Pérez, F. R. López-Estrada, G. Valencia-Palomo, L. Torres, V. Puig, and J. D. Mina-Antonio, "Leak diagnosis in pipelines using a combined artificial neural network approach," *Control Eng. Pract.*, vol. 107, Feb. 2021, Art. no. 104677.

[11] M.-U.-R.-A. Virk, M. F. Mysorewala, L. Cheded, and I. M. Ali, "Leak detection using flow-induced vibrations in pressurized wall-mounted water pipelines," *IEEE Access*, vol. 8, pp. 188673–188687, 2020.

[12] X. Zhou, Z. Tang, W. Xu, F. Meng, X. Chu, K. Xin, and G. Fu, "Deep learning identifies accurate burst locations in water distribution networks," *Water Res.*, vol. 166, Dec. 2019, Art. no. 115058.

[13] J. Mashford, D. De Silva, S. Burn, and D. Marney, "Leak detection in simulated water pipe networks using SVM," *Appl. Artif. Intell.*, vol. 26, no. 5, pp. 429–444, May 2012.

[14] Y. Liu, X. Ma, Y. Li, Y. Tie, Y. Zhang, and J. Gao, "Water pipeline leakage detection based on machine learning and wireless sensor networks," *Sensors*, vol. 19, no. 23, p. 5086, Nov. 2019.

[15] 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, Oct. 2016.

[16] A. Soldevila, R. M. Fernandez-Canti, J. Blesa, S. Tornil-Sin, and V. Puig, "Leak localization in water distribution networks using Bayesian classifiers," *J. Process Control*, vol. 55, pp. 1–9, Jul. 2017.

[17] 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. Comput. Sci.*, vol. 28, no. 2, pp. 283–295, Jun. 2018.

[18] C. Sun, B. Parellada, V. Puig, and G. Cembrano, "Leak localization in water distribution networks using pressure and data-driven classifier approach," *Water*, vol. 12, no. 1, p. 54, Dec. 2019.

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

[20] I. Lučin, B. Lučin, Z. Čarija, and A. Sikirica, "Data-driven leak localization in urban water distribution networks using big data for random forest classifier," *Mathematics*, vol. 9, no. 6, p. 672, Mar. 2021.

[21] W. Moczulski, R. Wyczółkowski, K. Ciupke, P. Przystałka, P. Tomasik, and D. Wachla, "A methodology of leakage detection and location in water distribution networks—The case study," in *Proc. 3rd Conf. Control Fault-Tolerant Syst. (SysTol)*, Sep. 2016, pp. 331–336.

[22] Q. Zhang, Z. Y. Wu, M. Zhao, J. Qi, Y. Huang, and H. Zhao, "Leakage zone identification in large-scale water distribution systems using multiclass support vector machines," *J. Water Resour. Planning Manage.*, vol. 142, no. 11, Nov. 2016, Art. no. 04016042.

[23] L. A. Rossman, "Epanet 2: Users manual," Nat. Risk Manage. Res. Lab., Office Res. Develop., U.S. Environ. Protection Agency, Cincinnati, OH, USA, Tech. Rep. EPA/600/R-00/057, 2000.

[24] *University of Exeter Centre for Water Systems*. Benchmarks. Accessed: Nov. 6, 2020. [Online]. Available: http://emps.exeter.ac.U.K./engineering/research/cws/downloads/benchmarks/

[25] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and M. Blondel, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[27] L. Grbčić, I. Lučin, L. Kranjčević, and S. Družeta, "Water supply network pollution source identification by random forest algorithm," *J. Hydroinformatics*, vol. 22, no. 6, pp. 1521–1535, Nov. 2020.

[28] J. P. Vítkovský, A. R. Simpson, and M. F. Lambert, "Leak detection and calibration using transients and genetic algorithms," *J. Water Resour. Planning Manage.*, vol. 126, no. 4, pp. 262–265, 2000.

[29] I. Lučin, L. Grbčić, S. Družeta, and Z. Čarija, "Source contamination detection using novel search space reduction coupled with optimization technique," *J. Water Resour. Planning Manage.*, vol. 147, no. 2, Feb. 2021, Art. no. 04020100.

[30] Z. Chi, Y. Li, W. Wang, C. Xu, and R. Yuan, "Detection of water pipeline leakage based on random forest," in *Proc. J. Phys., Conf.*, vol. 1978, 2021, Art. no. 012044.

[31] J. Cho, H. Kim, A. L. Gebreselassie, and D. Shin, "Deep neural network and random forest classifier for source tracking of chemical leaks using fence monitoring data," *J. Loss Prevention Process Ind.*, vol. 56, pp. 548–558, Nov. 2018.

[32] O. Akinsete and A. Oshingbesan, "Leak detection in natural gas pipelines using intelligent models," in *Proc. SPE Nigeria Annu. Int. Conf. Exhib.*, 2019, pp. 573–583.

[33] A. Sagheer, M. Zidan, and M. M. Abdelsamea, "A novel autonomous perceptron model for pattern classification applications," *Entropy*, vol. 21, no. 8, p. 763, Aug. 2019.

**IVANA LUČIN** received the B.S. and M.S. degrees in mechanical engineering from the Faculty of Engineering, University of Rijeka, in 2013 and 2015, respectively, where she is currently pursuing the Ph.D. degree in computational mechanics.

She is currently a Teaching Assistant with the Faculty of Engineering, University of Rijeka. Her research interests include hydraulic systems analysis, machine learning, and optimization methods.

**ZORAN ČARIJA** received the Ph.D. degree from the Faculty of Engineering, University of Rijeka, in 2007, with a dissertation in the field of computational mechanics.

He is currently a Professor with the Department of Fluid Mechanics and Computation Engineering, Faculty of Engineering, University of Rijeka. He is also the Head of the Section of Fluid Mechanics and Hydraulic Turbomachinery. His research interests include computational fluid dynamics, fluid mechanics, turbomachinery, water turbines, and renewable energy.

**SINIŠA DRUŽETA** received the Ph.D. degree from the Faculty of Engineering, University of Rijeka, in 2007, with a dissertation in the field of free surface flow modeling.

He is currently a Professor with the Department of Fluid Mechanics and Computation Engineering, Faculty of Engineering, University of Rijeka. He is the also the Head of the Section of Computational Engineering. His research interests include hydraulic systems analysis and optimization, open channel flow, and optimization methods.

**BOŽE LUČIN** received the B.S. and M.S. degrees in mechanical engineering from the Faculty of Engineering, University of Rijeka, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree in computational mechanics.

He is currently employed as a Project Engineer at Flowtech d.o.o., and he is also an External Associate with the Faculty of Engineering, University of Rijeka. His research interests include computational fluid dynamics and renewable energy.

• • •