Leveraging Transfer Learning in LSTM Neural Networks for Data-Efficient Burst Detection in Water Distribution Systems

Glynis, Konstantinos; Kapelan, Zoran; Bakker, Martijn; Taormina, Riccardo

DOI
10.1007/s11269-023-03637-3

Publication date
2023

Document Version
Final published version

Published in
Water Resources Management

Citation (APA)
Glynis, K., Kapelan, Z., Bakker, M., & Taormina, R. (2023). Leveraging Transfer Learning in LSTM Neural Networks for Data-Efficient Burst Detection in Water Distribution Systems. *Water Resources Management*, *37*(15), 5953-5972. https://doi.org/10.1007/s11269-023-03637-3

Water Resources Management
https://doi.org/10.1007/s11269-023-03637-3

Leveraging Transfer Learning in LSTM Neural Networks for Data-Efficient Burst Detection in Water Distribution Systems

Konstantinos Glynis[1,2] · **Zoran Kapelan**[1] · **Martijn Bakker**[2] · **Riccardo Taormina**[1]

Received: 27 March 2023 / Accepted: 2 October 2023

Abstract
Researchers and engineers employ machine learning (ML) tools to detect pipe bursts and prevent significant non-revenue water losses in water distribution systems (WDS). Nonetheless, many approaches developed so far consider a fixed number of sensors, which requires the ML model redevelopment and collection of sufficient data with the new sensor configuration for training. To overcome these issues, this study presents a novel approach based on Long Short-Term Memory neural networks (NNs) that leverages transfer learning to manage a varying number of sensors and retain good detection performance with limited training data. The proposed detection model first learns to reproduce the normal behavior of the system on a dataset obtained in burst-free conditions. The training process involves predicting flow and pressure one-time step ahead using historical data and time-related features as inputs. During testing, a post-prediction step flags potential bursts based on the comparison between the observations and model predictions using a time-varied error threshold. When adding new sensors, we implement transfer learning by replicating the weights of existing channels and then fine-tune the augmented NN. We evaluate the robustness of the methodology on simulated fire hydrant bursts and real-bursts in 10 district metered areas (DMAs) of the UK. For real bursts, we perform a sensitivity analysis to understand the impact of data resolution and error threshold on burst detection performance. The results obtained demonstrate that this ML-based methodology can achieve Precision of up to 98.1% in real-life settings and can identify bursts, even in data scarce conditions.

Keywords Deep learning · LSTM · Transfer learning · Burst detection · District metered areas

1 Introduction

Water distribution systems (WDSs) are underground networks designed to transport and distribute safe drinking water. Pipe bursts constitute a major challenge for WDS managers as they cause severe disturbance in the operation of the system, limit the availability

✉ Konstantinos Glynis
K.G.Glynis@tudelft.nl

[1] Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628 CN Delft, Netherlands

[2] Aquasuite® Research Department, Royal HaskoningDHV, P.O. Box 1132, 3800 BC Amersfoort, Netherlands

Published online: 12 October 2023

of sufficient and clean water (Al-washali et al. 2016; Fox et al. 2016), causing significant financial losses (Farley et al. 2001).

To reduce the impact of pipe bursts, water utilities resort to digitalization by installing pressure and flow monitoring sensors that automatically relay data to an operations center (Adedeji et al. 2017; Gupta and Kulat 2018). Monitoring allows water utilities to detect bursts early on, mobilize their repair crews swiftly and ultimately limit their negative consequences and promote economic and environmental sustainability (Cassidy et al. 2021; Bakker et al. 2012).

1.1 Related Studies

Timely detection of bursts is crucial to the water utilities, and two primary approaches exist: model- and data-driven approaches (Hu et al. 2021). Model-based approaches compare observations of the real network data with simulations of the WDS (Pérez et al. 2011). Despite numerous successful applications (Casillas Ponce et al. 2014; Sophocleous et al. 2019), these approaches require expert-calibrated models (Hu et al. 2021; Pérez et al. 2014) and a high degree of supervision by the user. Furthermore, these methods need expensive recalibration of the underlying hydraulic model when the WDS changes (Kang and Lansey 2011). On the other hand, data-driven methodologies rely on signal processing, statistical analysis, and machine learning (ML) to process the acquired data, disregarding in-depth understanding of the layout and operation of the WDS (Mounce et al. 2002). Recently, ML methods emerged as the most common data-driven approaches for burst detection. This family of methods usually work in a burst-no/burst binary classification fashion (Caputo and Pelagagge 2003; Mounce and Machell 2006; Mounce et al. 2014). However, acquiring balanced datasets for training is challenging since bursts are infrequent (Wu and Liu 2017). A proven strategy to tackle this issue involves initially training models to reproduce sensor trajectories on burst-free datasets. In the testing phase, the system identifies potential bursts by flagging deviations from the predicted values that surpass a set threshold (Hu et al. 2021; Romano et al. 2014).

All ML-based approaches for burst detection proposed in the literature operate with a fixed number of sensors or a set WDS topology. This requires the development of a new model every time there is a change in the sensor setup or in the physical network structure. Furthermore, training a new model relies on the acquisition of sufficient data under the new configuration, potentially leading to significant delays in detecting bursts. To overcome this issue, new models could reuse the knowledge captured by existing ones, rather than starting from scratch with each modification. This can be achieved by leveraging transfer learning, a ML technique that allows a model to apply knowledge learned from one task to a related task (Pan and Yang 2010; Torrey and Shavlik 2010). Unfortunately, common ML approaches based in non-parametric methods such as Decision Trees and Random Forest (Lučin et al. 2021; Zhang et al. 2022) do not transfer because they rely on fixed architectures that cannot adapt to new data distributions or changes in features. On the other hand, traditional Deep Learning (DL) architectures based on the Multi-Layer Perceptron (MLP) can be retrained to accommodate for changes in the data distributions of their inputs/outputs (e.g., due to a change in the physical network), but they cannot transfer knowledge when features are added or removed (i.e., following the installation or removal of sensors). Furthermore, MLPs are prone to suffer from the curse

of dimensionality, and require considerable amount of data to achieve good performances (Russell and Norvig 2010).

Modern DL methods like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are designed to bypass the curse of dimensionality by using inductive biases and shared parameters, which promote better knowledge transfer and smooth adaptation to varying data and input configurations (Bentivoglio et al. 2022). The sequential inductive bias of RNNs is particularly suitable for processing the time-series data measured by sensors in WDSs. Furthermore, gated RNN neurons, such as Long Short-Term Memory (LSTM) cells can effectively manage long sequences by selectively processing and propagating crucial information across time steps (Hochreiter and Schmidhuber 1996; Lai et al. 2018). This property renders them particularly attractive to handle the long-term correlations in flow, pressure and water demand data sensed in WDS. Despite these advantages, only a handful of studies have utilized LSTMs for burst detection. Wang et al. (2020) used an LSTM network and flow data to detect bursts in a real-life DMA in China, but their dataset was limited to simulated bursts and lacked pressure information. Similarly, Lee and Yoo (2021) worked with flow data to detect a single burst in a WDS, which is not representative of real DMAs. Xu et al. (2020) used flow and pressure signals with an LSTM but only tested on five fire hydrant simulated bursts in a non-DMA city-wide network.

No study explored how transfer learning in LSTMs can improve burst detection in operational settings. In this paper, we aim to address this gap by proposing a novel data efficient LSTM-based approach that leverages transferability to handle a varying number of sensors. When adding new sensors, we augment the original LSTM by duplicating weights for the newly added channels. These augmented models are fine-tuned, not re-trained from scratch, reducing data requirements for burst prediction under the modified setup. We validate our approach on simulated fire hydrant bursts and real bursts across 10 DMAs of Sutton and East Surrey Water Services Ltd (SES Water) in England. We also perform an extended sensitivity analysis to assess the impact of input time resolutions, providing insights into how data granularity affects the overall performance.

2 Case Studies

Table 1 reports the 10 anonymized DMAs of SES Water used in this study. Using satellite imagery, we identify three land use categories: urban, rural and mixed. Urban DMAs are characterized by dense urban fabric and very little unbuilt area. Rural DMAs are scarcely populated and are mostly covered by agricultural fields. Mixed DMAs lie in between the two previous categories. Regardless of their classification, all DMAs follow the layout depicted in Fig. 1 (left), with flow and pressure sensors installed at the inflow point, and an additional pressure sensor installed at the critical point. The three DMAs where fire hydrant bursts were simulated are listed in Table 2, and they have five to seven additional pressure sensors as shown in Fig. 1 (right). The data for this study was collected from 14 October 2016 to 29 March 2022, with varying data availability across the different sensors. All data has an original time resolution of 15-min.

The different length of the training, validation and testing subsets shown in Table 1 is a result of the requirement to have consistent flow and pressure signals, unaffected by sensor replacements and/or recalibration.

Table 1 Characteristics of DMAs used with training, validation and testing period partitioning

DMA	Land use	Subset*[a]	Period*[b]	Number of real bursts*[c]
Alpha	Urban	Training	1/10/2016 – 31/12/2018	11
		Validation		17
		Testing	1/1/2019 – 31/12/2020	41
Beta	Urban	Training	1/10/2016 – 31/12/2018	33
		Validation		19
		Testing	1/1/2019 – 31/12/2020	37
Gamma	Mixed	Training	1/1/2019 – 31/12/2020	28
		Validation		16
		Testing	1/1/2021 – 28/3/2022	21
Delta	Rural	Training	1/10/2016 – 31/5/2018	0
		Validation		1
		Testing	1/6/2018 – 31/12/2020	6
Epsilon	Urban	Training	1/10/2016 – 31/12/2018	35
		Validation		4
		Testing	1/1/2019 – 31/8/2020	60
Zeta	Mixed	Training	1/10/2016 – 31/12/2018	1
		Validation		6
		Testing	1/1/2019 – 31/5/2020	5
Eta	Rural	Training	1/10/2016 – 31/12/2018	6
		Validation		7
		Testing	1/1/2019 – 31/1/2020	8
Theta	Urban	Training	1/10/2016 – 31/5/2018	5
		Validation		4
		Testing	1/6/2018 – 28/2/2019	7
Iota	Rural	Training	1/4/2018 – 31/5/2020	15
		Validation		2
		Testing	1/6/2020 – 28/3/2022	4
Kappa	Mixed	Training	20/5/2021 – 31/12/2021	7
		Validation		0
		Testing	1/1/2022 – 28/3/2022	3

*[a]This subset partitioning refers to the model application on detecting real bursts

*[b]The training/validation partitioning follows the 70/30 ratio

*[c]Bursts in the training and validation subsets are removed, so that the model is trained in burst-free conditions

2.1 Simulated Fire Hydrant Bursts

Fire hydrant bursts were executed on March 10, 2022 (i.e., Beta and Delta DMAs) and March 15 (i.e., Zeta DMA) during daytime, after the installation of additional pressure sensors in early January 2022. The experiments were conducted for a total of 2.5 h with a progressively increasing discharge to avoid unnecessary harm to the network pipes. Details on the experiments, along with the burst discharge relative to the mean DMA inflow α_{burst}, are shown in Table 2.

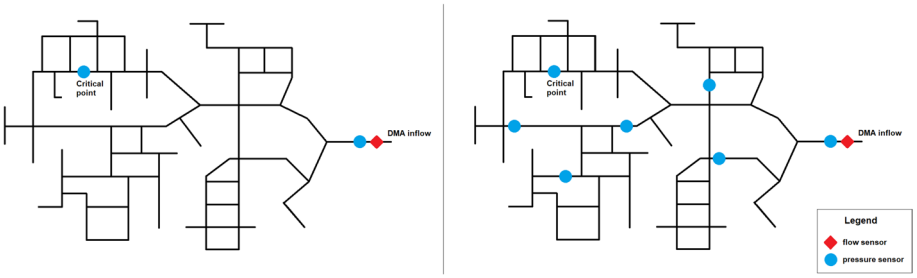

Fig. 1 A schematic representation of a typical DMA with distributed pressure and flow sensors for the case of real bursts (left) and simulated fire hydrant bursts (right)

At the original 15-min time resolution, each 2.5 h long simulated burst corresponds to 11 time steps, with the start and end time of the burst included.

2.2 Real Bursts

A total of 192 real bursts were available across the 10 DMAs (see Table 1). The burst records included detection datetime, repair datetime and a short description of their nature. As discussed later, the quality of measurements and the veracity of the burst records is inhomogeneous across the different DMAs. Furthermore, there are common challenges with the dataset. The registered bursts started before the operator detected them. As a result, the information available only partially represents "ground truth". Similarly, "burst-free" records contain background leaks and/or undetected bursts. Furthermore, it is possible that the sensors were recalibrated or replaced during the recording period, which undermines the consistency of the dataset.

3 Methodology

3.1 Overview of the Approach

The proposed detection mechanism works in a two-step prediction-classification fashion (Wu and Liu 2017). In the first step, the model makes a prediction of flow and pressure(s) for the next time step $t+1$, using autoregressive inputs until time $t-k$,

Table 2 Details on artificial fire hydrant experiments, including duration and burst size compared to mean DMA inflow

DMA	Beta		Delta		Zeta	
Fine-tuning*	16/1/2022 – 16/2/2022					
Testing	20/2/2022 – 28/3/2022					
Burst date	10/3/2022				15/3/2022	
$Q_{DMA,mean}$	7.2 l/s		1.3 l/s		2.5 l/s	
Duration	1.5 h	1 h	1.5 h	1 h	1.5 h	1 h
Q_{burst}	0.8 l/s	1.5 l/s	0.8 l/s	1.5 l/s	0.8 l/s	1.5 l/s
α_{burst}	11%	21%	62%	115%	32%	60%

*: Applicable to the scenarios that leverage transfer learning and fine-tuning of pre-trained weights

where k is a fixed time window, and known information of datetime-related features at time $t+1$. If H is the vector observed hydraulic features, the first stage can be expressed mathematically as

$$\widehat{H}_{t+1} = \varphi\left(H_t, H_{t-1}, \ldots, H_{t-k}, D_{t+1}\right) \tag{1}$$

where $\widehat{H}$ identifies the predicted hydraulic features, D are the datetime features and φ identifies the DL model. In the first stage, the goal of the DL model is to minimize the prediction error E_{t+1} in the training dataset without overfitting, expressed below for a single instance using the mean squared error or L_2 norm.

$$E_{t+1} = \frac{1}{n}\|H_{t+1} - \widehat{H}_{t+1}\|_2^2 \tag{2}$$

where n is the number of hydraulic features to predict. The choice of the squared error was mainly driven by the convexity of the metric, as well as the emphasis on larger errors, both of which simplify the optimization process in model training. In the second stage, bursts are flagged by comparing the prediction error against a time-varying threshold that changes with the time of the day to account for the cyclical nature of water demand (Hutton and Kapelan 2015). The thresholds are selected based on the distribution of the prediction errors on the validation dataset to strike a compromise between the sensitivity of the method and the excessive flagging of false positives (Taormina and Galelli 2018).

3.2 Input Features

In addition to past values of the hydraulic features H, we create two additional datetime features D, to help the DL model recognize the daily and weekly water consumption behavioral patterns (Hutton and Kapelan 2015). The first, named "Day Index" (DI) represents an engineered version of the weekday index with values in the [0, 1] range

$$DI = \frac{0.2}{1 - 0.8 \cdot \cos\left(((Weekday\ Index + 1)\%7) \cdot \frac{\pi}{3}\right)} \tag{3}$$

with Monday having a weekday index of 0, Sunday having a weekday index of 6 and "%" representing the modulo operator. According to Eq. (3), the DI values of working days are close to 0 and the DI values of weekends are equal to 1. The DI values of public holidays were set to 1 after statistical analysis confirmed behavioural patterns resembling weekend consumption, mainly due to the delayed morning peaks. The second datetime feature is the minute-of-the-day (MD), which accounts for the different expected consumption within the day and takes values in the range [0, 1439]. All hydraulic and datetime features are scaled in the range [0, 1] based on the value range they exhibit in the training dataset.

3.3 Neural Network Architecture

The proposed architecture includes both LSTM cells (Hochreiter & Schmidhuber 1997) and traditional neurons. Specifically, there are two different hidden layers; one consisting of 16 LSTM neurons that takes as input sequences of hydraulic features H and one

consisting of 2 regular neurons that takes as input singular values of the time features D. The number of LSTM neurons resulted from a preliminary hyperparameter tuning. The outputs of both layers are then concatenated into an additional layer consisting of regular neurons that predict the hydraulic features $\widehat{H}$. The number of neurons in the output layer is equal to the number of the predicted hydraulic features, i.e., 3 for real bursts and 8 (=3+5) or 10 (=3+7) for the engineered fire hydrant bursts. Regardless of the sensor setup, all models are trained to minimize the mean squared error in Eq. (2) computed for the entire training dataset. Finally, to reduce the possibility of overfitting, recurrent dropout is used when training the LSTM cells. Initial hyperparameter tuning revealed that of the different combinations of dropout rate and other parameters of the NN structure (details not provided here due to limited space), a dropout rate of 20% is preferred. This high rate is most likely justified by the relatively small size and noise of the dataset, especially for the simulated bursts.

3.4 Transfer Learning

Transfer learning refers to the "*improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned*" (Torrey and Shavlik 2010). This technique enables the development of a full-scale model when limited data is available for the new task, by building on an existing model that allows for knowledge transfer. In the burst detection domain, changes in the WDS topology, the number, calibration or type of sensors can significantly limit the length and consistency of datasets available for model training. Such changes often necessitate the model to be set "offline", until the assimilation of new, long-enough datasets are available and model re-training is possible. In this work, we leverage LSTM transferability for detecting the engineered fire hydrant bursts, since the additional pressure sensors are available only for the first three months of 2022. Specifically, we first develop a model using the data before 2022. Second, we expand this model by adding new input channels to the LSTM, corresponding to the number of new pressure sensors. Then, we initialize the additional trainable parameters in the LSTM gates with the weights from the pressure sensor at the inflow of the DMA. This results in a new model with pre-trained weights for all pressure signals, which is then fine-tuned and tested with limited data to detect the fire hydrant bursts.

3.5 Multi-Threshold Classification

The second stage of the proposed approach requires the comparison of the prediction error against set thresholds. These thresholds are set at the 99.9^{th} percentile of the prediction errors observed in the validation dataset. This value was selected to limit the false positives to 0.1% in operational settings, assuming the ideal case where the validation dataset accurately represents test conditions. However, due to the periodical components of water demand patterns, the variation of the prediction error is usually higher during periods of intensely varying water consumption, and smaller when water consumption is relatively stable, e.g., during night-time (Hutton and Kapelan 2015). To account for the heterogeneity in the error distribution, we segmented the prediction errors into $h = 3$ hour intervals, creating contiguous clusters. Moreover, we

distinguished between working days and weekends or public holidays. This resulted in 16 distinct thresholds, each based on the 99.9-th percentile of prediction error distributions within their respective clusters. This number of clusters offered the best trade-off between the threshold resolution and the reliability of the computed percentiles. Smaller values of h created too small clusters from which to extract a high percentile. This impacted the model performance by "flagging" too many non-burst instances as bursts. Higher values of h resulted instead in a too coarse clustering of the daily consumption behavior and loss of fidelity.

3.6 Performance Assessment

We employed both event- and timestamp-based metrics to assess the methodology. The event-based metrics factor in unique alarms — instances where prediction errors exceed the time-dependent threshold — that occur within the repair timeframe and one week before a burst is identified by the operator. This one-week lead time accounts for the potential delays in burst detection, which can be due to delayed customer reports or the time it takes to recognize substantial water loss. Conversely, we derived timestamp-based metrics on a per value basis by comparing the generated alarm instances to the burst records. Event-based metrics are useful to water utilities and engineers, since they show the performance in terms of number of bursts, whereas the value-based ones are more useful to NN experts, since they reflect the per-datapoint performance of the model.

Three event-based metrics are calculated: $Recall_e$, $Precision_e$ and $f1score_e$, where subscript "e" denotes the event-based calculation of these metrics. The timestamp-based metrics calculated are *Recall*, *Fallout* and *Precision*, defined as shown in Table 3.

Bearing in mind the distinctions between event- and timestamp-based metrics, *TP* in Table 3 represent true positives, which are correct flags raised for actual instances (either events or timestamps); *FN* denotes false negatives, where true instances are not flagged; *TN* refers to true negatives, cases where no burst occurs and no flag is raised; and *FP* indicates false positives, instances where a flag is raised erroneously as no burst is occurring.

Based on the above definitions, the *Recall* metrics indicate the proportion of actual burst events that were correctly identified, emphasizing the model's sensitivity to detecting bursts. On the other hand, *Precision* represents the proportion of flagged burst events that are true burst events, showcasing the reliability of the alarms generated. The $f1score$ is the harmonic mean of Precision and Recall, a composite measure that is suited for imbalanced datasets, such as the one in this study where burst-free conditions are the norm. Lastly, *Fallout* is the probability of false alarms. A good methodology should have high *Recall*, *Precision* and $f1score$, whereas *Fallout* should be minimal. For simulated burst cases, we introduce an additional *Detection Delay* (*DD*) metric. This metric measures the time interval between the actual onset of a fire hydrant burst and the moment the burst is flagged.

Table 3 Event- and timestamp-based performance metrics

Event-based metrics	Timestamp-based metrics
$Recall_e = \frac{TP}{TP+FN} \times 100\%$	$Recall = \frac{TP}{TP+FN} \times 100\%$
$Precision_e = \frac{TP}{TP+FP} \times 100\%$	$Precision = \frac{TP}{TP+FP} \times 100\%$
$f1score_e = \frac{2 \cdot (Recall_e \cdot Precision_e)}{Recall_e + Precision_e} \times 100\%$	$Fallout = \frac{FP}{FP+TN} \times 100\%$

3.7 Experimental Setup

We split the datasets for each DMA into two parts: one for training and validation, that is cleaned of any registered bursts, and one for testing. To prevent the occurrence of vanishing gradients in detecting real bursts, we reduce the length of the input sequence fed to the LSTM cells by resampling the original data at 30-min resolution. This yields input sequences for the hydraulic variables of 192 timesteps, with data up to 4 days before the prediction time. We implement the neural networks using the Keras package in Python. We train all models for 100 epochs using the Adam optimizer (Kingma and Ba 2015), storing the best model based on the validation performance at each epoch. The optimizer's suitability of this problem stems from its adaptive learning rate mechanism, its ability to store moving averages of both the gradients and the squared gradients of the parameters, and not their past instances. The model is trained using a decaying learning rate and early stopping, regulated on the validation loss. The initial learning is 0.010 and reduced by 20% whenever the validation loss does not significantly for 6 consecutive epochs. Similarly, the training stops if the validation loss does not improve significantly after 10 epochs. For both cases, a loss improvement of 0.001 is used as tolerance.

4 Results and Discussion

4.1 Results for Fire Hydrant Bursts

Three different scenarios are initially investigated: (i) Scenario A, where only the 3 originally installed sensors are used, but the models are developed using all the data available for training and validation (see Table 1), (ii) Scenario B, where all the available sensors are utilized, but the models are developed only with data from 2022, with training and validation taking place in the "fine-tuning" period of early 2022 (see Table 2), (iii) Scenario C, with all the available sensors, but with transfer learning of the model weights from scenario A, and fine-tuning the expanded model for a period of 1 month (see Table 2).

Table 4 shows that for scenario A, only 1 of the 3 performed fire hydrant bursts is detected and only after a delay of 60-min. This is explained by the distant location of the bursts compared to the original pressure sensors and the significant fragmentation of the training and validation subsets. Because Scenario A leverages the longest training and validation subsets compared to the other scenarios, it is plagued by multiple sensor replacements/recalibration and operator-induced pressure adjustments, which limited the "usable" dataset length to be fed to the LSTM neurons. Scenario B results are also poor due to the very short length of the training subset with all sensors, which does not capture the inter-annual seasonality.

The benefits of training on data before adding additional sensors and then employing transfer learning become clearer when we compare scenarios B and C; the latter of which does use transfer learning. We see improvements in performance across the board, especially in the fact that we can detect all three bursts within either 15 min or 30 min.

Regardless of the considered Scenario, *Fallout* values for the DMA Delta are high. Upon closer investigation of the raw time-series, it was found that the pressure sensor installed at the critical point was faulty, with pressure readings increasing by more than 30 m. To assess the impact of the faulty sensor, Table 4 reports two additional

Table 4 Detection performance on fire hydrant bursts

DMA	Number of sensors		Bursts		Performance metrics (timestamp-based)			
	Q	P	Performed	Detected	*Recall*	*Fallout*	*Precision*	*DD*(min)
Scenario A: Additional sensors not utilized								
Beta	1	2	1	0	0.0%	1.0%	0.0%	x
Delta	1	2	1	0	0.0%	23.0%	0.0%	x
Zeta	1	2	1	1	9.1%	0.8%	5.3%	60
Scenario B: Additional sensors utilized without transfer-learning								
Beta	1	2+7	1	0	0.0%	5.6%	0.0%	X
Delta	1	2+5	1	1	81.8%	23.0%	1.8%	30
Zeta	1	2+5	1	1	9.1%	2.2%	6.0%	60
Scenario C: Additional sensors utilized with transfer-learning								
Beta	**1**	**2+7**	**1**	**1**	**9.1%**	**2.7%**	**1.9%**	**15**
Delta	1	2+5	1	1	90.9%	22.6%	1.9%	30
Zeta	**1**	**2+5**	**1**	**1**	**36.4%**	**2.8%**	**14.3%**	**30**
Scenario D: Additional sensors utilized without transfer-learning and P_{CP} removed								
Beta	1	1+7	1	0	0.0%	6.1%	0.0%	X
Delta	1	1+5	1	1	45.5%	2.1%	9.6%	30
Zeta	1	1+5	1	0	0.0%	3.1%	0.0%	x
Scenario E: Additional sensors utilized with transfer-learning and P_{CP} removed								
Beta	1	1+7	1	1	9.1%	3.3%	1.3%	15
Delta	**1**	**1+5**	**1**	**1**	**27.3%**	**1.7%**	**7.0%**	**30**
Zeta	1	1+5	1	1	18.2%	4.4%	1.9%	60

Values in bold font relate to the optimal information input stream configuration for each DMA. For DMAs Beta and Zeta that involves inclusion of the additional pressure sensors with transfer learning. For DMA Delta the removal of a pressure sensor is also included

scenarios: Scenarios D and E, which are identical to B and C respectively, with the removal of the pressure sensor at the critical point.

As expected, the performance on DMA Delta improves for both Scenarios, with *Fallout* plummeting from 23.0% and 22.6% to 2.1% and 1.7%, respectively. Also *Precision* increases from 1.8% and 1.9% to 9.6% and 7.0%, respectively. For the same DMA *DD* is not improved and *Recall* is reduced. The latter is caused by the misclassification of a few burst instances as false negatives. However, this has limited to no impact in the actual detection, since *DD* remains unchanged to 30 min.

For DMAs Beta and Zeta that had no faulty sensor, the exclusion of the sensor at the critical point deteriorates the performance in scenarios D and E compared to B and C respectively. This is reflected on the *Recall* decrease, the *Fallout* increase, and *Precision* decrease. Especially for DMA Zeta, scenario D parameter selection is detrimental and leads to the burst going completely unnoticed (with a *Recall* of 0%). This signifies the importance of the spatial coverage of every sensor and that unnecessary information removal has consequences for the overall performance.

The robustness of the method in detecting the simulated bursts is supported by the fact that the bursts in DMAs Beta and Zeta for Scenario C and DMA Delta for Scenario E (where the problematic pressure sensor signal was removed, takes place within either one, i.e., 15-min, or two, i.e., 30-min, time steps. Furthermore, for these cases, fall out is 2.8%, indicating a very low rate of false alarms.

Figure 2 presents a 24-h snapshot of the fire hydrant bursts, originated from an identical, yet different run of the algorithm. A visual inspection of these figures reveals the existence of residual alarms, i.e., additional exceedances of the error threshold, after the simulated burst stops. This phenomenon is likely due to the lingering effects of the recent burst; the LSTM cells continue to use the disrupted hydraulic data from the burst period in their subsequent predictions for some time. Although operators can readily eliminate these incorrect flags following a burst repair, we have chosen to categorize them as false alarms in this study. This classification yields higher *Fallout* and lower *Precision*.

Fig. 2 Fire hydrant bursts in DMAs Beta (top), Delta (middle) and Zeta (bottom). Left sub-figures are for Scenario A. Right sub-figures are for Scenario C. Top row is for the discharge at the inflow of the DMA.. Middle row is for the pressure at the inflow of the DMA and the critical point. Lower row is for the MSE (error), the error threshold, the burst start and repair time and the raised alarms

4.2 Detection of Real Bursts

Table 5 shows that real burst detection performance varies greatly. $Recall_e$ $Precision_e$ $f1score_e$ range from very low to very high values. Similar ranges are exhibited in the timestamp-based metrics. Performance in DMA Epsilon is very good, with the highest $f1score_e$ of 66.7%, the lowest *fallout* of 0.2% and the highest *Precision* of 98.%. Performances are particularly low for DMA Delta, with a $f1score_e$ of 6.7%, the highest *fallout* of 12.4% and the lowest *Precision* of 12.2%.

It is also worth noting the high correlation between the number of bursts recorded in each DMA and the $Precision_e$ metric, evidenced by a correlation coefficient of 0.848. A substantial correlation is also observed between the timestamp-based *Precision* aand the number of bursts, with a coefficient of 0.750. Given that the datasets for the various DMAs are roughly equal in length (see Table 1), it is plausible that these correlations arise from varying degrees of public alertness, which typically plays a significant role in pipe burst identification.

The unreported bursts may explain the differences in detection performance across the DMAs investigated. This claim is supported by the land use cover of the different DMAs. Namely, DMAs Delta and Eta that exhibit the worse performance are mostly rural, whereas DMAs Alpha, Beta and especially Epsilon are heavily urbanized. This is an indication that several actual bursts in the rural DMAs may go completely unnoticed. This has a two-fold impact on our methodology. First, not all bursts are removed for training and validation, thus impairing the ability of the model to "learn" burst-free (i.e., normal) behavior. Second, the existence of multiple unregistered bursts in the testing subset leads to an overwhelming number of *FP* s, which should be in fact be labeled as *TPs*.

4.3 Sensitivity Analysis

To study the effect of time resolution on the burst detection performance, the entire process of training, validation and testing is repeated for the urban DMA Beta and the rural DMA

Table 5 Performance of LSTM-based model on real bursts. For the detection of the real bursts a single flow sensor and two pressure sensors were utilized, as described in the Section 2

DMA	Number of bursts	Event-based Metrics			Timestamp-based Metrics		
		$Recall_e$	$Precision_e$	$f1score_e$	*Recall*	*Fallout*	*Precision*
Alpha	41	29.3%	63.2%	40.0%	6.3%	0.5%	86.9%
Beta	37	29.7%	78.6%	43.1%	0.6%	0.1%	96.9%
Gamma	21	38.1%	47.1%	42.1%	9.1%	4.7%	57.1%
Delta	6	16.7%	4.2%	6.7%	16.4%	12.4%	12.2%
Epsilon	60	68.3%	65.1%	66.7%	10.6%	3.2%	98.1%
Zeta	5	40.0%	28.6%	33.3%	5.4%	0.2%	51.4%
Eta	8	62.5%	3.0%	5.7%	8.9%	3.0%	61.7%
Theta	7	57.1%	26.7%	36.4%	15.5%	2.7%	79.4%
Iota	4	50.0%	22.2%	30.8%	7.2%	0.5%	40.9%
Kappa	3	100.0%	23.1%	37.5%	6.3%	1.4%	57.2%

Eta. Different combinations of time resolution (15 min, 30 min and 60 min) and length of the input hydraulic feature time-series (1 to 7 days) are investigated. Tow conditions apply to this analysis: (1) The length of the time-series is limited to a maximum of 250, so that the efficiency of the LSTM cells is not severely reduced by vanishing gradients; (2) An integer number of days is used, so as to not interfere with the 24-h behavior seasonality of water consumption. Table 6 shows the event- and timestamp-based performance metrics for the two DMAs along with the original combination of length = 4 days and time resolution of 30 min used in the previous experiments.

Table 6 shows that the impact of time resolution and/or length of the input time-series is not negligible. More specifically, coarser resolution leads to significantly higher values of the *Fallout* and lower *Precision*., which translates to less confidence on the alarms. This phenomenon can be attributed to the relative prominence of spurious exceedances of the error thresholds when compared to the same number of corresponding instances in finer resolution datasets.

Furthermore, in the 60-min resolution, the model cannot "decode" the short-term dynamics of bursts, because pressure and flow data are aggregated into 1-h intervals. Even though there is a slight increase in the values of both $Recall_e$ and *Recall* for coarser resolution, it is maybe preferable to sacrifice the detection of a handful of bursts, for the sake of higher confidence, which is reflected in the increase of all the other metrics. Based on these results, it emerges that combination of 2-day long time-series at 15-min resolution is superior with respect to all the other combinations. This is supported by the better scores across all the performance metrics. However, a coarser resolution of 30 min provides more time to the sensors to relay their measurements.

Furthermore, we studied the sensitivity of the performance to the time-varying error threshold. This analysis was performed only for DMA Beta, which is the most heavily urbanized and it is characterized by the highest number of registered bursts. The results are shown in Fig. 3 for the metrics $Recall_e$, $f1score_e$, $Precision_e$ and *Precision*.

As is expected, Fig. 3 shows that lower thresholds lead to higher sensitivity, with more bursts being detected and an overall higher $Recall_e$. However, this is accompanied by more false alarms impacting *Precision*, which reduces significantly from over 80% to less than 50%. This trade-off is also seen in $f1score_e$, which has the largest value for the 99.9^{th} percentile, for both curves. As for *Precision*, lowering the error threshold reduces it, but not as much as $Precision_e$.

4.4 Comparison to Other Burst Detection Methods

We assess our burst detection method through qualitative comparison with existing LSTM-based techniques. The lack of available code implementation and the difference in the case studies prevents direct comparisons. Wang et al. (2020) employed a pure LSTM model to detect both simulated and synthetic bursts in a single DMA. In detecting simulated bursts, their model was able to detect bursts after two time steps using 5-min resolution data. This is comparable to our findings (see Table 4) although we were able to detect bursts also at the first time-step and at coarser resolutions. Lee and Yoo (2021) implemented a different LSTM model with flow data only, to detect a single burst. Their approach yielded inferior results (sensor-based $Recall \in (46.46\%, 99.81\%)$ and $Fallout \in (0.11\%, 29.88\%)$ compared to $Recall_e \in (16.7\%, 100\%)$ and $Fallout \in (0.10\%, 12.4\%)$. In summary, our burst detection

Table 6 Sensitivity analysis of input time-series length and time resolution

DMA		Beta						Eta					
Time resolution [min]		60	30	15	60	30	15	60	30	15	60	30	15
Metric		*Recall$_e$*			*Recall*			*Recall$_e$*			*Recall*		
Length of LSTM-fed time-series [days]	7	*35.1%*			*1.9%*			50.0%			5.4%		
	6	*37.8%*			*1.6%*			**62.5%**			7.6%		
	5	*29.7%*	27.0%		*1.5%*	0.3%		**75.0%**	*62.5%*		*14.1%*	*10.8%*	
	4	*37.8%*	**29.7%**		*1.4%*	**0.6%**		50.0%	**62.5%**		*9.7%*	**8.9%**	
	3	*35.1%*	24.3%		*1.3%*	0.1%		50.0%	*75.0%*		4.9%	*13.5%*	
	2	*37.8%*	*32.4%*	*37.8%*	*1.3%*	0.5%	*1.7%*	*62.5%*	*75.0%*	*75.0%*	*12.4%*	*22.4%*	*18.7%*
	1	*37.8%*	*29.7%*	27.0%	*1.0%*	0.4%	0.2%	*62.5%*	*62.5%*	*75.0%*	*10.5%*	8.8%	*19.2%*
Metric		*Precision$_e$*			*Precision*			*Precision$_e$*			*Precision*		
Length of LSTM-fed time-series [days]	7	76.5%			*97.3%*			*5.8%*			55.1%		
	6	66.7%			94.0%			*6.0%*			56.6%		
	5	61.1%	71.4%		94.4%	92.3%		*4.0%*	2.9%		50.4%	*62.6%*	
	4	63.6%	**78.6%**		93.4%	**96.9%**		*4.4%*	**3.0%**		*63.5%*	**61.7%**	
	3	65.0%	69.3%		93.4%	84.0%		*6.6%*	2.6%		52.5%	60.2%	
	2	73.7%	70.6%	*93.3%*	95.1%	94.4%	*98.9%*	*3.9%*	1.7%	*10.3%*	53.1%	53.2%	*65.1%*
	1	70.0%	57.9%	*83.3%*	93.0%	87.7%	*97.3%*	*3.7%*	*3.1%*	*8.8%*	54.7%	60.8%	*68.7%*
Metric		*f*1*score$_e$*			*Fallout*			*f*1*score$_e$*			*Fallout*		
Length of LSTM-fed time-series [days]	7	*48.2%*			*0.1%*			*10.4%*			*2.4%*		
	6	*48.3%*			0.2%			*10.9%*			3.1%		
	5	40.0%	39.2%		0.2%	*0.0%*		*7.5%*	5.5%		7.5%	3.5%	
	4	*47.4%*	**43.1%**		0.2%	**0.1%**		*8.0%*	**5.7%**		*3.0%*	**3.0%**	
	3	*45.6%*	36.0%		0.2%	*0.0%*		*11.6%*	5.0%		*2.4%*	4.8%	
	2	*50.0%*	*44.4%*	*53.8%*	0.1%	0.1%	*0.0%*	*7.3%*	3.4%	*18.2%*	5.9%	10.7%	*2.7%*
	1	*49.1%*	39.3%	40.8%	0.2%	0.1%	*0.0%*	*7.0%*	*5.8%*	*15.8%*	4.7%	3.1%	*2.5%*

Bold represents the combination of time-series length used in the previous experiments. Italic and underline represent better and worse performance respectively

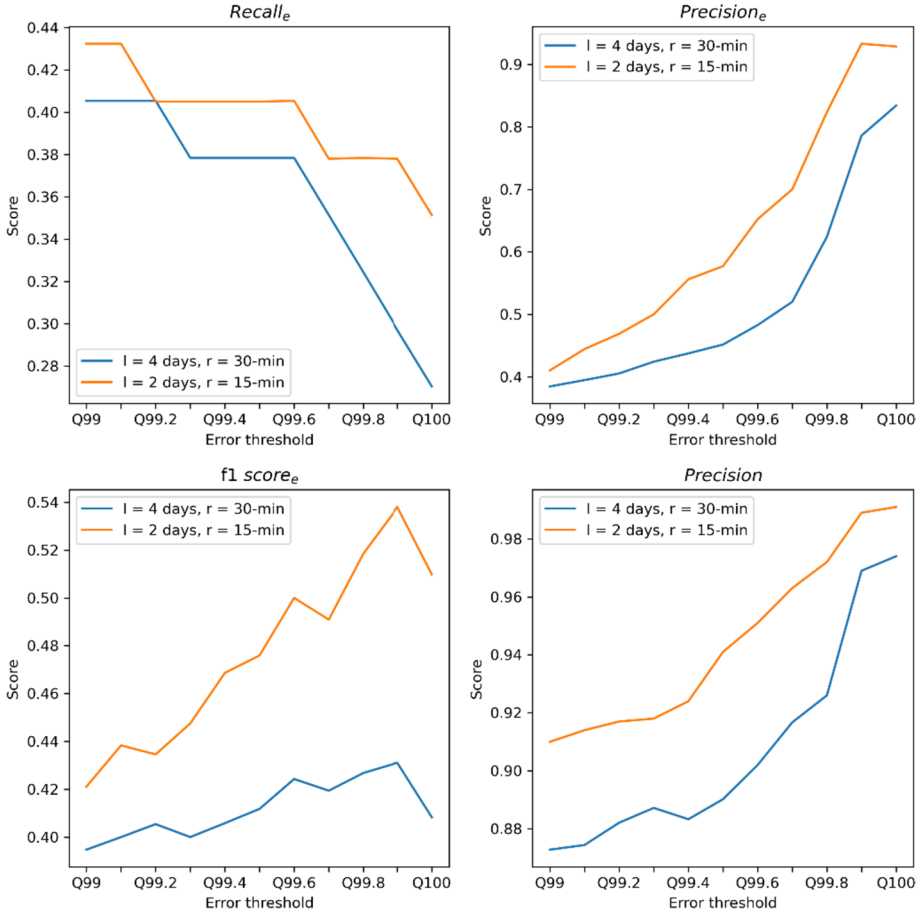

Fig. 3 Sensitivity of the performance metrics to the error threshold

method exhibits promise compared to existing LSTM methods, with specific performance variations dependent on the dataset and methodology employed.

We conducted a quantitative comparison against Autoencoders (AE), a DL architecture successfully used in the detection of other types of anomalies in WDS, i.e., cyber-physical attacks (Taormina and Galelli 2018). We use AEs to compare their performance against our LSTM-based method on the same real burst dataset. As demonstrated in Table 7 of the Appendix, our LSTM model outperforms the AE approach, as there is an overall improvement in $f1$ *score* computed across all the DMAs. This highlights the LSTM effectiveness in detecting bursts, most likely due to the incorporation of sequential inductive bias.

Evaluation of the model's performance on well-established benchmark datasets does not take place. Even though this would facilitate comparisons with existing methodologies, it would not suffice for direct comparisons with established LSTM-based

techniques. None of the existing LSTM methodologies has undertaken such comparisons, and they hold particular significance for our approach as we aim to enhance upon them. Moreover, assessing the adaptability of previously benchmarked models is challenging due to their structural constraints, which often preclude the incorporation of inductive bias and handling extremely brief datasets, as exemplified by the two-month records for the simulated bursts in this context.

5 Conclusions

This work presents a novel LSTM-based method for pipe burst detection in water distribution systems. The developed model harnesses the potential of LSTM networks to predict flow and pressure during normal operational circumstances. Notably, the algorithm exhibits elevated prediction errors when exposed to data stemming from pipe burst incidents. The salient attributes of the LSTM architecture encompass its power in managing extensive temporal sequences and its inherent adaptability that enables integration and exclusion of information streams. Importantly, the technique leverages transfer learning to overcome the constraints arising from limited training datasets and a varying number of sensors.

Testing on real bursts in 10 DMAs in the UK revealed that the developed LSTM-based method exhibits varying performance, with $Recall_e \in [16.7\%, 100\%]$ and $Fallout \in [0.1\%, 12.4\%]$ reflecting a varying confidence to the correct identification of bursts and the incorrect classification of non-bursts respectively. This inconsistent performance across DMAs was noted due to the, sometimes poor, burst record quality. This data is crucial for training the model in burst-free conditions and correlating alarms to actual bursts. Limited public awareness, especially in rural DMAs, and unnoticeable smaller bursts impact the proposed approach. Finer data resolution, namely 15-min time steps, enables capturing abrupt discontinuities, such as pipe bursts better, and enhances burst-detection performance, For urban settings this increases $Precision_e$ from 78.6% to 93.3% and for rural settings the same metric increases from 3.0% to 10.3%. Sensitivity analysis shows that variable error threshold mirrored to the daily water consumption behavior further improves burst detection robustness.

Testing on simulated fire hydrant bursts emphasizes burst-sensor proximity. As noted before in the literature, distant bursts pose detection challenges. However, installation of additional sensors in the DMA reduces this issue and enables the timely detection of bursts corresponding to as little as 11% of the mean DMA inflow. LSTM's inherent flexibility and transfer learning facilitate easy integration of extra data streams.

It is paramount to acknowledge the study's inherent limitations, notably the requisite reliance on burst-free training pressure and flow datasets. The acquisition of such datasets remains an arduous task, particularly within sparsely populated rural settings or DMAs characterized by a scarcity of monitoring sensors. The LSTM-based burst detection algorithm sensitivity to sensor recalibration and replacement accentuates the necessity for meticulous identification of temporally consistent measurement periods suitable for model training and validation. In addition, the alarm persistence past the burst repair is noted as a weakness too and can be circumvented, at an operational level, by temporarily "suspending" the model from running after alarms are raised. Finally, testing this methodology in simulated bursts taking place at nighttime is recommended for evaluating its robustness in more realistic conditions.

Appendix

Table 7 Performance of Autoencoder on real bursts

DMA	Sensors		Bursts	Event-based Metrics			Timestamp-based Metrics		
	Q	P		$Recall_e$	$Precision_e$	$f1score_e$	*Recall*	*Fallout*	*Precision*
Alpha	1	2	41	14.6%	30.0%	19.7%	0.6%	0.5%	54.0%
Beta	1	2	37	32.5%	41.7%	36.5%	0.3%	0.6%	48.5%
Gamma	1	2	21	28.6%	7.4%	11.8%	5.5%	7.4%	27.6%
Delta	1	2	6	0.0%	0.0%	0.0%	0.0%	0.6%	0.0%
Epsilon	1	2	60	26.7%	94.1%	41.6%	2.9%	0.5%	99.4%
Zeta	1	2	5	16.9%	3.9%	6.3%	2.0%	2.2%	14.3%
Eta	1	2	8	37.5%	13.6%	20.0%	6.2%	2.8%	54.8%
Theta	1	2	7	28.6%	11.8%	16.7%	4.3%	4.2%	35.9%
Iota	1	2	4	25.0%	100.0%	40.0%	0.4%	0.0%	100.0%
Kappa	1	2	3	66.7%	5.3%	9.8%	6.8%	19.9%	15.3%

Details of Comparison to the Autoencoder

The AEs feature an encoder-decoder architecture that learns a compressed representation of high-dimensional input data via error minimization. The abstraction of information and the lack of any explicit hydraulic information of the system make AE very robust in detecting anomalies in real-life settings. However, a significant difference with our problem is that the AE used by Taormina and Galelli (2018) was working on 43 different sensors at the same time; whereas here we are using it on way less sensors, but many self-correlated inputs.

The AE-based burst detection method developed here uses the AE fed with a $n = 4$-day long time-series of the flow and pressure signals from the DMA inflow, pressure at the critical point and the same two time features as before: Day Index and minute-of-the-day. Due to the self-supervised nature of the AE, the goal of training is recreating the input as output with the highest fidelity possible. The Adam optimizer (Kingma and Ba 2015) is used again, as is the case with the LSTM-based model, and the ReLu function is employed to activate the neurons. Hyperparameter tuning is very important for the AE as it is in the case of LSTM networks. For this reason, preliminary investigation was carried out to determine what is the best possible combination of the AE layers and neurons that enables a reasonably deep understanding of the network dynamics without having an overwhelmingly large number of trainable weights. An Autoencoder with successive layers of 64, 32, 16, 32 and 64 neurons all resulted in the optimal structure for this problem. The results of applying the AE on real bursts analyzed in this paper are provided in Table 7.

As it can be seen from Table 7, the Autoencoder fails to capture the behaviour dynamics of the specific real-life DMAs in focus, as the burst detection performance is significantly inferior to the one of the LSTM-based model (see Table 5). All the metrics, both event- and timestamp-based imply that the AE is not sensitive enough to understand the discrepancies in the system caused by pipe bursts. The only seemingly improved metrics are the lower (i.e. better) values of *Fallout* for DMAs Delta, Epsilon, Eta and Iota. However, the fact

that these are so low, almost approaching zero, in combination with the also low values of *Recall* imply that the AE prediction error remains below the set threshold for most of the time-series length.

Acknowledgements This study was conducted as part of a MSc thesis and graduate internship project between the Faculty of Civil Engineering and Geosciences of Delft University of Technology and Royal HaskoningDHV. The authors would like to thank Elvin Isufi from the Department of Intelligent Systems of Faculty of Electrical Engineering, Mathematics and Computer Science of TU Delft and Haochen Zhang from Royal HaskoningDHV for their advice and insight into the operation of water distribution systems and machine learning. The authors would also like to thank the Sutton and East Surrey Water Services Ltd (SES Water, Redhill, UK), for providing all the data used in this work.

Author Contributions All authors contributed to the study conception and design. K. Glynis performed material preparation, data collection, software development and analysis. K. Glynis wrote the original draft of the manuscript. R. Taormina wrote parts of the original draft. R. Taormina, Z. Kapelan, and M. Bakker reviewed and edited the paper. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Availability of Data and Materials The Python code scripts created for this study can be provided upon request.

Declarations

Ethical Approval The authors subscribe to the ethical principles of this journal.

Consent to Participate All authors approved to participate in the efforts to publish this paper.

Consent to Publish All authors approve the publication in this journal.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Adedeji K, Hamam Y, Abe B, Abu-Mahfouz A (2017) Towards achieving a reliable leakage detection and localization algorithm for application in water piping networks: An overview. IEEE Access 5:20272–20285. https://doi.org/10.1109/ACCESS.2017.2752802

Al-washali T, Sharma S, Kennedy M (2016) Methods of assessment of water losses in water supply systems: a review. Water Resour Manag 30:4985–5001. https://doi.org/10.1007/s11269-016-1503-7

Bakker M, Vreeburg JH, Rietveld LC, Van De Roer M (2012) Reducing customer minutes lost by anomaly detection?. In WDSA 2012: 14th Water Distribution Systems Analysis Conference, 24-27 September 2012 in Adelaide, South Australia. Barton, ACT: Engineers Australia, pp 913–927. https://search.informit.org/doi/10.3316/informit.946749511368491

Bentivoglio R, Isufi E, Jonkman SN, Taormina R (2022) Deep learning methods for flood mapping: a review of existing applications and future research directions. Hydrol Earth Syst Sci 26(16):4345–4378. https://doi.org/10.5194/hess-26-4345-2022

Caputo AC, Pelagagge PM (2003) Using neural networks to monitor piping systems. Process Saf Prog 22(2):119–127. https://doi.org/10.1002/prs.680220208

Casillas Ponce MV, Garza Castanon LE, Cayuela VP (2014) Model-based leak detection and location in water distribution networks considering an extended-horizon analysis of pressure sensitivities. J Hydroinf 16(3):649–670. https://doi.org/10.2166/hydro.2013.019

Cassidy J, Barbosa B, Damião M, Ramalho P, Ganhão A, Santos A, Feliciano J (2021) Taking water efficiency to the next level: digital tools to reduce non-revenue water. J Hydroinf 23(3):453–465. https://doi.org/10.2166/hydro.2020.072

Farley M, Water S, Supply W, Council SC, WHO WH (2001) Leakage management and control: a best practice training manual. Farley, Malcolm, No. WHO/SDE/WSH/01.1. World Health Organization, 2001. World Health Organization. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/66893/WHO_SDE_WSH_01.1_eng.pdf

Fox S, Shepherd W, Collins R, Boxall J (2016) Experimental quantification of contaminant ingress into a buried leaking pipe during transient events. J Hydraul Eng 142(1):04015036. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001040

Gupta A, Kulat KD (2018) A selective literature review on leak management techniques for water distribution system. Water Resour Manag 32:3247–3269. https://doi.org/10.1007/s11269-018-1985-6

Hochreiter S, Schmidhuber J (1996) LSTM can solve hard long time lag problems. Adv Neural Inf Process Syst 9:473–479. Retrieved from https://proceedings.neurips.cc/paper/1996/hash/a4d2f0d23dcc84ce983ff9157f8b7f88-Abstract.html

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu Z, Chen B, Chen W, Tan D, Shen D (2021) Review of model-based and data-driven approaches for leak detection and location in water distribution systems. Water Supply 21(7):3282–3306. https://doi.org/10.2166/ws.2021.101

Hutton C, Kapelan Z (2015) A probabilistic methodology for quantifying, diagnosing and reducing model structural and predictive errors in short term water demand forecasting. Environ Model Softw 66:87–97. https://doi.org/10.1016/j.envsoft.2014.12.021

Kang D, Lansey K (2011) Demand and roughness estimation in water distribution systems. J Water Resour Plan Manag 137(1):20–30. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000086

Kingma D, Ba J (2015) Adam: A method for stochastic optimization. International Conference on Learning Representations. San Diego, California, United States: ICLR 2015. https://doi.org/10.48550/arXiv.1412.6980

Lai G, Chang WC, Yang Y, Liu H (2018) Modeling long-and short-term temporal patterns with deep neural networks. 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 95–104). Ann Arbor, MI, USA: ACM. https://doi.org/10.1145/3209978.3210006

Lee C, Yoo D (2021) Development of Leakage Detection Model and Its Application for Water Distribution Networks Using RNN-LSTM. Sustainability 13(16):9262. https://doi.org/10.3390/su13169262

Lučin I, Lučin B, Čarija Z, Sikirica A (2021) Data-driven leak localization in urban water distribution networks using big data for random forest classifier. Mathematics 9(6):672. https://doi.org/10.3390/math9060672

Mounce SR, Machell J (2006) Burst detection using hydraulic data from water distribution systems with artificial neural networks. Urban Water J 3(1):21–31. https://doi.org/10.1080/15730620600578538

Mounce SR, Mounce RB, Jackson T, Austin J, Boxall JB (2014) Pattern matching and associative artificial neural networks for water distribution system time series data analysis. J Hydroinf 16(3):617–632. https://doi.org/10.2166/hydro.2013.057

Mounce S, Day A, Wood A, Khan A, Widdop P, Machell J (2002) A neural network approach to burst detection. Water Sci Technol 45(4–5):237–246. https://doi.org/10.2166/wst.2002.0595

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359. https://doi.org/10.1109/TKDE.2009.191

Pérez R, Cugueró M, Cugueró J, Sanz G (2014) Accuracy assessment of leak localisation method depending on available measurements. Procedia Eng 70:1304–1313. https://doi.org/10.1016/j.proeng.2014.02.144

Pérez R, Puig V, Pascual J, Quevedo J, Landeros E, Peralta A (2011) Methodology for leakage isolation using pressure sensitivity analysis in water distribution networks. Control Eng Pract 19(10):1157–1167. https://doi.org/10.1016/j.conengprac.2011.06.004

Romano M, Kapelan Z, Savić D (2014) Automated detection of pipe bursts and other events in water distribution systems. J Water Resour Plan Manag 140(4):457–467. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000339

Russell SJ, Norvig P (2010) Artificial Intelligence: A Modern Approach. (3rd ed.). Upper Saddle River, NJ: Pearson Education. ISBN-13: 978–0–13–604259–4

Sophocleous S, Savić D, Kapelan Z (2019) Leak localization in a real water distribution network based on search-space reduction. J Water Resour Plan Manag 145(7):04019024. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001079

Taormina R, Galelli S (2018) Deep-learning approach to the detection and localization of cyber-physical attacks on water distribution systems. J Water Resour Plan Manag 144(10):04018065. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000983

Torrey L, Shavlik J (2010) Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI global, pp 242–264. https://doi.org/10.4018/978-1-60566-766-9.ch011

Wang X, Guo G, Liu S, Wu Y, Xu X, Smith K (2020) Burst detection in district metering areas using deep learning method. J Water Resour Plan Manag 146(6):04020031. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001223

Wu Y, Liu S (2017) A review of data-driven approaches for burst detection in water distribution systems. Urban Water J 14(9):972–983. https://doi.org/10.1080/1573062X.2017.1279191

Xu Z, Ying Z, Li Y, He B, Chen Y (2020) Pressure prediction and abnormal working conditions detection of water supply network based on LSTM. Water Supply 20(3):963–974. https://doi.org/10.2166/ws.2020.013

Zhang X, Long Z, Yao T, Zhou H, Yu T, Zhou Y (2022) Real-time burst detection based on multiple features of pressure data. Water Supply 22(2):1474–1491. https://doi.org/10.2166/ws.2021.337