TU Delft

Development of an LSTM-based methodology for burst detection in water distribution systems

Thesis report

Konstantinos Glynis

20 June 2022

Development of an LSTM-based methodology for burst detection in water distribution systems

Konstantinos Glynis[1, 3]; Zoran Kapelan[1]; Elvin Isufi[2], Martijn Bakker[3] and Riccardo Taormina[1]

Abstract: Water utilities face many challenges, including pipe bursts that cause significant non-revenue water losses. Detecting those bursts early is important for the water sector in its path to achieve sustainable water resource management. This study presents a scalable data-driven methodology for burst detection in water distribution systems that is based on Long Short-Term Memory (LSTM)-based neural networks (NNs) and includes two stages: prediction and classification. Time-series of hydraulic (flow and pressure) signals are fed to the LSTM, whereas domain (time) features of the next time step are fed independently to regular neurons. These two streams of information are then concatenated to predict the values of the hydraulic features of the next time step. The model is trained on normal conditions only, so that when fed with data corresponding to a burst, the predictions will mismatch the observations. Comparison of the predictions to the observations is quantified though an error function, which is then used for classification. Specifically, a variable error threshold that corresponds to a pre-defined extreme percentile of the error distribution is used to discern bursts from normal conditions. The methodology is corroborated on two different types of bursts: (a) real bursts in district metered areas (DMAs) in the United Kingdom and (b) simulated fire hydrant leak tests in the same DMAs. For the real bursts, sensitivity analysis of the algorithm is performed to assess how data resolution and error threshold affect the performance. The flexibility of the method is studied for the simulated fire hydrant leaks, where additional information streams from new sensors are incorporated in the model by means of applying transfer learning and fine-tuning. The results obtained demonstrate that this scalable LSTM-based methodology works reasonably well in real-life settings and can successfully identify burst events, both real and simulated, even in DMAs with a small number of installed sensors. Furthermore, it is assessed how the flexibility of the LSTM neurons is pivotal for burst detection when utilizing a varying number of sensors.

Keywords: Deep learning; LSTM, Transfer learning; Burst detection; District metered areas

[1]: Faculty of Civil Engineering and Geosciences, TU Delft, Delft, Netherlands

[2]: Faculty of Electrical Engineering, Mathematics and Computer Science, TU Delft, Delft, Netherlands

[3]: Aquasuite® Research Department, Royal HaskoningDHV, Amersfoort, Netherlands

1. Introduction

Water distribution systems (WDSs) are underground networks designed to transport and distribute safe drinking water to a populated area, either urban or rural. They are comprised of numerous sections, interconnections, joints, valves, meters and fire hydrants. Water utility companies that operate these networks face a variety of challenges including hard-to-predict water demand, limited availability of suitable water resources and infrastructure obsolescence. Pipe bursts constitute one such challenge, as they cause severe disturbance in the operation of the system, the availability of sufficient and/or clean water (Fox et al., 2016), as well as financial losses (Farley et al. 2001).

To reduce the impact of the problem of pipe bursts, the water utility sector is progressively transforming the control and operation of water distribution systems by installing pressure and flow monitoring sensors that automatically relay data to an operations center (Adedeji et al., 2017). With this network-centric monitoring approach, water utilities can use the sensor data 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). Burst duration is characterized by the unawareness, awareness, localization and repair periods (Bakker et al., 2012; Mounce & Boxall, 2010). Large bursts are usually reported by clients or are easily identifiable from the magnitude of the flow/pressure deviation from the usual patterns. Smaller bursts or bursts that take place during night time and/or at remote locations, however, may stay unnoticed (Bakker et al., 2014). This leads to an extended period of unawareness with negative consequences.

Limiting this first period of burst unawareness and detecting these pipe failures quickly is important to the water utilities, explaining the vast amount of research in this field. Various techniques have already been applied by researchers, both model-based and data-driven ones (Hu et al., 2021). Model-based approaches are based on comparing observations of the real-life network with results of simulations of a digital parallel of the water distribution system (Brdys & Ulanicki, 1996; Pérez et al., 2011). Data-driven methodologies are based on signal processing or statistical analysis of the acquired data, that do not require an in-depth understanding of either the layout or the operation of the water distribution network (Mounce et al., 2002).

Model-based methodologies have shown great potential in burst detection (Casillas Ponce et al., 2014; Sophocleous et al., 2019). However, their requirement for acquiring large amount of data for calibrating the hydraulic parameters of the model increases their computational complexity and makes them more difficult to use (Pérez et al., 2014). In addition, if the topological features of the water distribution system change for any reason, e.g., subsidence, the hydraulic model of this type of approach requires reconstruction and/or recalibration (Kang & Lansey, 2011). This task is difficult for water utility companies to perform, but rather requires the inclusion of outside-of-the-organization experts (Hu et al., 2021). This ultimately increases the cost of using model-based methodologies in the long run, while also requiring a high degree of supervision by the user.

Data-driven methodologies, because of their lack of knowledge of explicit hydraulic principles, are not as prone to changes in the topological features or the hydraulic parameters of the network pipes as the model-based approaches, but are susceptible to insufficient or erroneous monitoring data (Romano et al., 2014). Thus, data availability plays a major role in their performance, especially because they have no knowledge of the physical network structure. Furthermore, the burst-no burst classification problem is usually an imbalanced one, which creates additional problems in the accuracy of data-driven methodologies (Oliker & Ostfeld, 2014). It is also worth mentioning that to explain the variability of the monitored parameters caused by factors other than bursts, e.g., end-user behaviour, input features external to the pipe network (e.g., time) can be used, to improve prediction accuracy of these methods (Ye & Fenner, 2014).

The majority of the developed data-driven methodologies, let alone the model-based ones, have been formulated with a specific type of DMA in mind, either real-life or simulated. Consequently, their robustness in effectively detecting bursts has been assessed in either real or simulated bursts. Therefore, there is a gap in transcending the boundary between detecting real and simulated bursts. Furthermore, most approaches have a fixed number of DMA installed sensors in mind, which for the purposes of training and testing is invariable. Hence, in addition to the application universality of existing burst detection methodologies, there is also room for improvement in making these approaches easily scalable in terms of including additional sensor signals without compromising their trainability and computational efficiency.

This scalability is defining the gap this research wants to fill. Namely, the development of a data-driven methodology for burst detection utilizing hydraulic (namely pressure and flow) and domain (time) features that is flexible and can integrate information from a varying number of monitoring sensors, while being parsimonious and keeping a small number of trainable parameters.

The innovative features of this research are three. First, using a Deep Learning architecture that combines Long Short-Term Memory (LSTM) cells for processing time-series of past hydraulic features and regular neurons for processing time attributes of the next time step for which a prediction of the hydraulic features takes place. Second, assessing the sensitivity of the methodology on different time resolutions of input data, which provides insight into the effect data granularity has in the detection confidence. And third, utilizing transfer learning to carry the pre-trained knowledge from a subset of nodal connections into new ones, in order to integrate information from additional sensors under conditions of data frugality.

The paper is organized as follows. After Section 1, i.e., the Introduction, follows Section 2, where related studies in the field of burst detection are presented. After that, Section 3 provides relevant information regarding the case studies and how they can be considered representative of water distribution systems. In Section 4, the theoretical background and data analyses of the devised methodology are presented. Then in Section 5 the results obtained from applying this methodology are presented and discussed. Finally, Section 6 provides conclusions deduced from the application of this approach recommendations for future work are given.

2. Related study

Artificial neural networks (ANNs) have been extensively used at the core of data-driven leak detection methodologies. This research domain has actually evolved rapidly since first developing the concept of feed-forward neural networks. Simple leak-no leak classification methods using ANNs have been used to detect pipe failures in both water distribution networks and piping networks carrying hazardous fluids (Caputo & Pelagagge, 2003; Mounce & Machell, 2006; Mounce et al., 2014; Sun et al., 2020). These methods, however, require extensive datasets with a good balance with and without leaks, which are not always so easy to acquire (Wu & Liu, 2017). Other approaches that use ANNs incorporate a prediction stage before classifying the data as exhibiting leaks or not (Mounce et al., 2003; Romano et al., 2011). At the prediction stage, data to be expected under normal conditions are estimated and their deviation to the observations is used to classify them (Hu et al., 2021).

Recurrent neural networks (RNNs) constitute a subcategory of ANNs that includes feedback (closed loop) connections (Fausett, 1994) and because they are ideally suited for treating long time-series, they have been widely adopted in the field of time-series prediction (Lai et al., 2018; Yu et al., 2017). However, in cases of very long time-series, having RNNs with gradient based learning and backpropagation is dangerous because of the problem of vanishing or exploding gradient (Pascanu et al., 2013). This has been solved to a great extent though gated architectures such as LSTMs (Cho et al., 2014; Hochreiter & Schmidhuber, 1997).

In burst detection applications there has been limited use of LSTM architectures so far, even though they have shown great potential in time-series forecasting applications (Siami-Namini et al., 2018). Wang et al. (2020) worked with an LSTM network and flow data only to detect several simulated bursts in a real-life DMA in south China. Although their method outperformed conventional approaches, it was tested on short datasets containing only simulated bursts and did not include any pressure information. Considering that pressure sensors are usually greater in number than flow sensors, this significantly limits the breadth of the hydraulic signatures the model can comprehend, let alone capturing real pipe bursts. Lee & Yoo (2021) worked with flow data only as well and tested their model on one burst of a water distribution network main line, which is not representative of real-world DMAs. Xu et al. (2020) worked with both flow and pressure signals as input to an LSTM-based model to predict pressure. This case study was based on detecting bursts on a whole city-wide, non-DMA partitioned water distribution system covering more than 100 km^2 in area but was tested on only five fire hydrant simulated leak tests utilizing 17 pressure sensors. Since the proliferation of partitioning a WDS into distinct DMA units however, it is dubious how such a wide spatial focus enables burst detection at the operational level in real-life settings. Bjerke (2019) also worked with LSTM cells and used pressure data from two academic (simulated) water distribution networks to detect simulated bursts.

Hence, it is understandable that not enough research has been done on LSTM-based methodologies for burst detection in typical DMA units and their potential in urban drinking water time-series forecasting have not been investigated adequately. Furthermore, the sensitivity of such methods to the time resolution of input data has not been assessed, as well as their adaptability to incorporate additional sensed data when the monitoring conditions in the network change.

3. Case studies

This paper focuses on applying a data-driven burst detection methodology on several real-life district metered areas (DMAs) in the United Kingdom for which frequent measurements of flow and pressure exist. The reason why it is decided to apply our model at the DMA scale of a water distribution network stems from the proliferation of this mode of spatial partitioning of a network (Morrison et al., 2007), which renders working at a WDS scale for burst detection moot.

These DMAs consist of an enclosed network of pipes and junctions controlled by a valve at the intake. Fig. (1) shows a typical DMA configuration, with flow and pressure sensors installed at the inflow point. An additional pressure sensor is installed at the critical point, e.g., the junction usually experiencing the lowest pressure in the entire DMA. The DMAs where fire hydrant leak tests took place have a similar layout as the others, but with the addition of five to seven more pressure sensors, scattered throughout the DMA. This allows for better capturing the network dynamics when bursts take place.

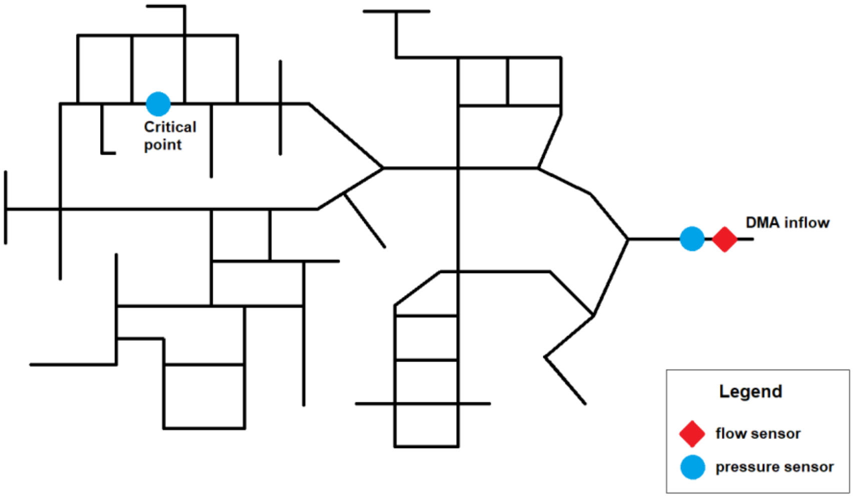

Figure 1: A schematic representation of a typical DMA. Pressure and flow sensors are installed at the inflow point, whereas an additional pressure sensor is installed at the critical point.

4. Methodology

4.1 Problem definition

Let us consider that burst detection takes place using both flow Q and pressure P measurements that come to the operational center at regular intervals (e.g., every 15-min) from sensors installed in the DMA and that a pipe burst occurs at a time between time steps $t_{start} - 1$ and t_{start}. Then the first measurements of both monitored features that will include the effect of the pipe burst will be the ones at time step t_{start}. Hence, flow $Q(t),\ t <$

t_{start} and pressure measurements $P(t),\ t < t_{start}$ will correspond to normal operating conditions. If we also assume that the burst is repaired at time t_{repair}, then $Q(t)$ and $P(t)$ for $t \in [t_{start}, t_{repair}]$ will not correspond to normal conditions, but to a burst.

The model developed here works in a prediction-classification fashion. For the specific case, the model makes a prediction of flow and pressure for the next time step t, based on combination of past information of these features spanning a period of days $[t - \Delta t, t - 1]$, where Δt is a fixed time window, and information coming from the domain (time) features for the next time step t. If we denote the prediction of feature X at time step t as $X'(t)$, the goal is for: $P'(t) = P(t)$ and $Q'(t) = Q(t)$ when there is no burst ($t < t_{start}$ and $t > t_{repair}$), but $P'(t) \neq P(t)$ and $Q'(t) \neq Q(t)$ for when there is a burst, $t \in [t_{start}, t_{repair}]$.

However, since $P'(t) = P(t)$ and $Q'(t) = Q(t)$ is almost unachievable, a threshold of dissimilarity is introduced, below which a prediction is labeled as corresponding to normal flow conditions and above which is labeled as exhibiting a burst. If this measure of dissimilarity is denoted as an error function E and its threshold above which a burst is detected as E_{thres}, then if $E(P(t), P'(t), Q(t), Q'(t)) \geq E_{thres}$ conditions exhibit a burst, whereas if $E(P(t), P'(t), Q(t), Q'(t)) < E_{thres}$ the conditions are normal.

This error function E, can be any one of the error functions used in the bibliography, including, but not limited to: mean absolute error, mean squared error, mean square log error and coefficient of determination.

The goal is for the model to have a good performance, meaning that the error function exceeds the threshold E_{thres} starting from the very first erroneous values at time $t = t_{start}$, and drops back below it once the burst has been repaired at time t_{repair}. A simplified illustration of how burst detection through the use of the error function of the created model works is shown in Fig. (2).

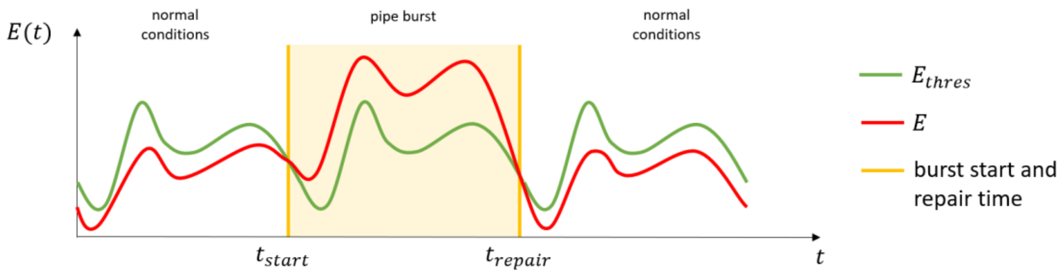

Figure 2: Detection of bursts through exceedance of error function threshold

4.2 General approach

Detection of bursts takes place through the use of the error function that quantifies the deviation of the predicted to the observed values of the hydraulic features. For the predictions to be close to the observations under normal operating conditions and not exceed the threshold, leak-free periods need to be extracted, on which the model is to be trained. Thus, the general approach here is training the model on data that do not exhibit leaks, so that the model learns the nominal behaviour of the network. If the model is trained well, the prediction error shall remain below the threshold for new leak-free data, but exceed it during bursts.

The first step includes acquiring the data of the monitored hydraulic (flow and pressure) variables of the network, as well as the leak job records, indicating the start and repair date and time of all the leaks jobs that took place in the DMAs. The leak job records include bursts, but also (scheduled) repairs of faulty elements of the network, such as valves or meters. Since the time features utilized here are engineered, we first extract them from the timestamps of the registered data. Then, part of the dataset is kept completely unseen and is separated for testing purposes. The remaining part of the dataset undergoes a process of registered leak jobs removal, so that it only contains leak-free data on which training and validation takes place.

Training and validation take place before testing the model. The LSTM-based model is trained to minimize the error function of the testing subset, while also taking care to not overfit by monitoring the progress of the error function reduction in the validation subset. In the next step, knowing that the error distribution is skewed with a lower bound of zero (no negative error exists), a statistically significant, e.g., 99.9-th percentile, value of the error function from the validation subset is extracted and is used as the threshold for detecting bursts. Testing makes use of the trained model and the not-previously seen part of the dataset that describes both normal conditions and bursts. In the last step, performance assessment takes place by comparing the timing of the registered leak jobs and the exceedances of the error threshold as this was extracted from the validation subset. Fig. (3) shows a schematic overview of the methodology, along with the basic flow of information between the steps described.

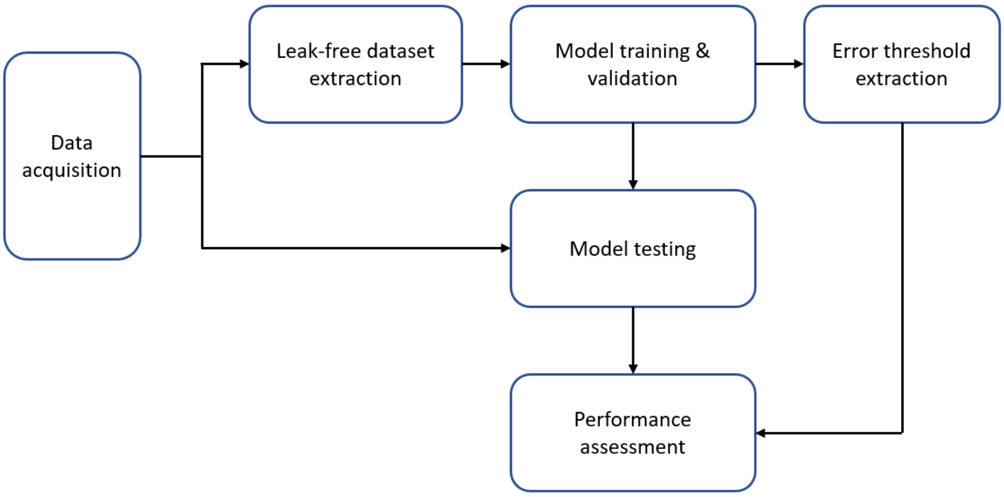
Data acquisition
Leak-free dataset extraction
Model training & validation
Error threshold extraction
Model testing
Performance assessment

Figure 3: Schematic representation of the methodology

4.3 Feature extraction

The model has information about the state of the network (flow and pressure) available, as well as the time each variable is recorded. The two different categories of features are used in a different way within the LSTM-based model. Specifically, n-day long time-series of hydraulic features at a resolution r that cover the entire n-days up to including the previous time step are used in combination with the domain (time) features of the next, i.e., current, time step to predict the hydraulic features of the next, i.e., current, time step.

If we denote H a hydraulic feature and T a time feature, we use the time-series $H_{t-(n/r)}, \ldots, H_{t-1}$ and the value of T_t to predict the value H_t. As explained later, multiple hydraulic and time features are used, but the principle is the same. The actual number of hydraulic features used actually depends on the type of burst and dataset that is used at each time. For detecting real leak jobs three hydraulic features are used, whereas for artificial fire hydrant leak tests this is expanded to include information from additionally installed pressure sensors.

For detecting real leak jobs, the flow and pressure measurements from the intake of the DMA are used in combination with the pressure measurement at the critical point of the DMA. For the artificial fire hydrant leak tests, in addition to the regular flow and pressure sensors at the inflow of the DMA and the pressure sensor at the critical point, there are also available five or seven (depending on the DMA) additional pressure sensors scattered throughout the DMA. So, two different sets of hydraulic features are examined in this case, one where only the three hydraulic features are used, as is the case with real bursts, and one where all the additional pressure signals are included. The raw hydraulic features are used as input to the LSTM part of the model because such cells are very well suited in understanding the implicit relationships and dynamics of the features, without the need of engineering them.

We also use two time related features. The first one, named "Day Index." (D.I.) represents an engineered version of the weekday index that takes values in the range [0, 1], in a way that resembles the typical end user behavior within the week. If we assume that weekdays start from Monday (with weekday index of 0) and end up on Sunday (with weekday index of 6) and "%" represents the remainder of division, then according to Eq. (1), the D.I. values of working days are close to 0 and the D.I. values of weekends are equal to 1. In addition, special care has been taken for public holidays, where a weekday index of 6 corresponding to Sundays has been assigned, so as to compensate for the different end user consumption patterns during such days. Investigative analysis at the beginning of this work showed that consumption behavior during such public holidays resembles the weekend consumption, mainly due to the delayed morning peak

$$D.I. = \frac{0.2}{1 - 0.8 \cdot \left(cos\left((Weekday\ index + 1)\ \%\ 7 \right) \cdot \frac{\pi}{3} \right)} \qquad (1)$$

The second time related feature is the minute-of-the-day (M.D.) that takes values in the range [0, 1440). This feature was used so as to account for the different expected consumption within the day.

All features are normalized, i.e., scaled, in the range [0, 1] based on the value range they exhibit in the training phase. Such a process is necessary so that the nodal connections of the neural network have weights of the same order of magnitude, which makes overfitting and unstable training less likely.

4.4 Neural network

The neural network used in this study is LSTM-based. Thus, it makes use of both regular and LSTM neurons. Regular neurons have an input-output process embedded in them and are the building blocks of traditional artificial neural networks (ANNs). For the formal description of regular neurons, the so-called mapping function Γ is used. The mapping function assigns for each neuron j a subset $\Gamma(j) \subseteq V$ which consists of all ancestors of the current neuron (Svozil et al., 1997). Each neuron in a particular layer is connected with multiple (in many cases with all the) neurons of the previous layer. The connection between the j-th and the k-th neuron is characterized by a weight coefficient ω_{jk} and the j-th neuron by a threshold θ_j. The weight coefficient ω depends on the degree of importance of a connection to the degree the ANN provides a good approximation of the desirable output. The output value of the j-th neuron y_j is determined by Eqs. (2)-(3):

$$y_j = f(\xi_j) \qquad (2)$$

$$\xi_j = \theta_j + \sum_{k \epsilon \Gamma_j^{-1}} \omega_{jk} y_k \qquad (3)$$

where ξ_j is the potential of the j-th neuron and $f(\xi_j)$ is the so-called transfer function. Note that in Eq. (3) summation takes place over all neurons k transferring information to the j-th neuron. As for the threshold coefficient, it can be understood as a weight coefficient of the connection with formally added neuron k, where the output value $y_k = 1$ (bias). For the transfer function, it is known that:

$$f(\xi) = \frac{1}{1 + e^{-\xi}} \qquad (4)$$

Recurrent neural networks (RNNs) are different to ANNs, because they use self-connections from the previous time step as inputs. Thus, the hidden state of the RNN neurons contains a dynamic history of the input features sequence instead of a fixed-size window (Schmidhuber, 2015). RNNs are notoriously difficult to be trained on long time-series because they exhibit the problem of vanishing/exploding gradient. However, recent advances in the field and the introduction of the gated LSTM architecture have greatly improved their ability to process nonlinear data, and especially long sequences of it.

Compared to conventional hidden units, the use of LSTM cells ensures that when training, during the stage of back-propagation, the gradient does not vanish or increase sharply (usually referred to as "explosion") after a large number of iterations. Fig. (4) illustrates the elements that constitute an LSTM unit.

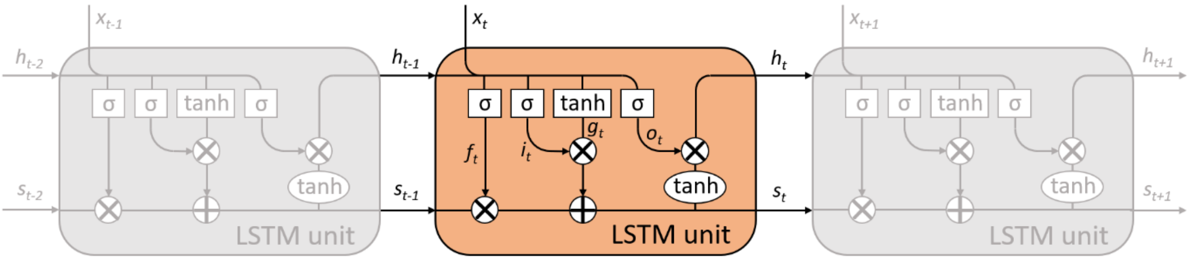

Figure 4: Illustration of an LSTM unit

The input sequence of LSTM is $x = \{x(1),\ x(2), \dots ,\ x(n/r)\}$, where $x(t)$ is an H-dimensional vector in this study, with H the number of hydraulic features, n the number of consecutive rolling days extracted from the dataset and r the resolution of the dataset. The calculation process of an LSTM cell can be briefly explained by Eqs. (5)-(10):

$$g_t = tanh\left(W^{xg}x_t + W^{hg}h_{t-1} + b_g\right) \qquad (5)$$

$$i_t = \sigma\left(W^{xi}x_t + W^{hi}h_{t-1} + b_i\right) \qquad (6)$$

$$f_t = \sigma\left(W^{xf}x_t + W^{hf}h_{t-1} + b_f\right) \qquad (7)$$

$$s_t = s_{t-1} \odot f_t + i_t \odot g_t \qquad (8)$$

$$o_t = \sigma\left(W^{xo}x_t + W^{ho}h_{t-1} + b_o\right) \qquad (9)$$

$$h_t = o_t \odot tanh(s_t) \qquad (10)$$

where h_t is the output; s_t is the cell state; $\odot$ denotes element multiplication; g_t, i_t, f_t and o_t represent the squeeze unit, input unit, forget unit and output unit respectively; σ denotes sigmoid function; W^{xg}, W^{hg}, W^{xi}, W^{hi}, W^{xf}, W^{hf}, W^{xo} and W^{ho} are related weight matrices; and b_g, b_i, b_f and b_o are related biases.

In this study, the neural network has a composite structure including both LSTM and regular neurons. Specifically, there are two different input layers; one consisting of 16 LSTM neurons that takes as input sequences of hydraulic features and one consisting of 2 regular neurons that takes as input singular values of the domain (time) features. The number of the LSTM neurons resulted from an initial hyperparameter tuning, since either less or more neurons led to poorer performance. The outputs of both layers are then concatenated into an output layer consisting of regular neurons that give as output the values of the hydraulic features for the time step defined by the domain (time) features fed. The number of neurons in the output layer is equal to the number of the (predicted) hydraulic features. For real leaks it is 3, and for simulated fire hydrant leaks it is 3 for the case of no additional pressure sensors used and 3+5 or 3+7 for the cases of including the additional pressure sensors. The reason for choosing this specific structure of the neural network is the fact that LSTM cells are better suited for processing long time-series, whereas regular neurons are more easily trainable and have adequate complexity for processing singular values of the time-step corresponding to the prediction. Also, a dropout rate of 20% in the LSTM cells is used to limit the possibility of overfitting. A schematic representation of the LSTM-based neural network is shown in Fig. (5).

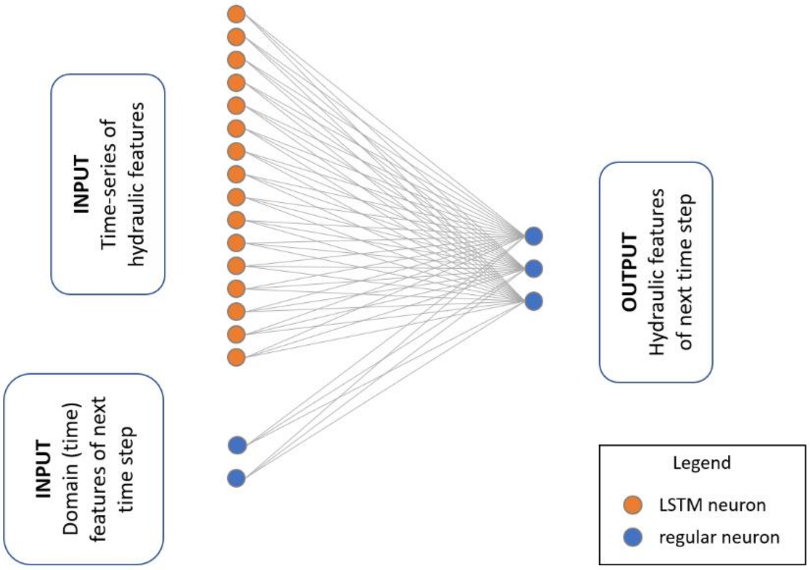
INPUT
Time-series of hydraulic features
INPUT
Domain (time) features of next time step
OUTPUT
Hydraulic features of next time step
Legend
● LSTM neuron
● regular neuron

Figure 5: Typical structure of the neural network for the case of using three (3) hydraulic features

4.5 Transfer learning

Transfer learning consists of "*taking features learned on one problem, and leveraging them on a new, similar problem"* (Keras, 2020). For instance, features from a model that has learned to identify primates in image collections may be useful to kick start another model aimed at identifying humans in image collections. Transfer learning usually takes place under two conditions. First, a dataset is too short for training a full-scale model from scratch. Second, the model is flexible by its nature and enables its weights to be transferred/adjusted with minimal impact on its trainability.

Both of these conditions apply to the case of testing our LSTM-based model on the simulated fire hydrant leak tests. Specifically, in the DMAs where the tests took place, the time-period from that point when additional pressure sensors are installed until the time when the fire hydrant leak tests take place is really short, i.e., just two and a half months, for the model to be trained, validated and tested properly. For this reason, the model is initially implemented on the time period prior to the installation of the additional sensors with the features used before (the three hydraulic and two time features). Then the knowledge (weights) of the nodal connections corresponding to the pressure time-series from the inflow of the DMA are replicated as the nodal connections of the additional pressure time-series to be utilized. In that way, a new model with pre-trained weights for all the pressure signals, even the ones that did not exist prior to the installation of the pressure sensors, is created. The model is then fine-tuned and then applied on the fire hydrant leak tests.

4.6 Multithreshold classification

The LSTM-based model is applied to predict the hydraulic features under normal flow conditions. When a burst occurs, the abnormal values of the observed hydraulic features differ from the predictions. This can be quantified by the error function E, which in this study corresponds to the squared error. The formula of the error function for a time step t is shown in Eq. (11):

$$E_t = \sum \left(H_t - \widehat{H}_t\right)^2 \qquad (11)$$

where H_t is a vector of the observed hydraulic features and $\widehat{H}_t$ a vector of the predicted hydraulic features. According to Eq. (11) the error is the sum of the deviations of the predictions of all the hydraulic features to their observations, which is possible because of the prior normalization of all the features. Initial analysis of a customized weighted average of the error function that put more weight to flow compared to pressure did not yield the expected results. Hence, the pressure signals that are often plagued by "benching" and operator-induced displacements proved to be more reliable in estimating the error function.

Since the nominal end-used behaviour in a water distribution system is dynamic and varies within the 24 h/day, so is the behaviour of the hydraulic features monitored as well as the behaviour of the residual (squared) errors (Hutton & Kapelan, 2015). This means that the variation of the prediction error is naturally higher during periods of intensely varying water consumption, e.g., during daytime, whereas during periods of relatively stable water consumption, e.g., during night-time, the prediction error is smaller. If a fixed threshold is applied throughout the day, then bursts during night-time will go unnoticed, whereas this will likely yield many false alarms during daytime. The same differential behaviour of the system dynamics is observed between working days and weekends/public holidays.

To account for this heterogeneity of the error distribution, the prediction (squared) errors are split into different clusters, one for each h-hour interval and different for working days and weekends/public holidays. Multiple values of h were initially examined, and $h = 3$ was finally selected, since it offered the best trade-off between diurnal error threshold resolution and a large enough dataset from which to extract very high percentile values. Hence, if the number of predictions within one day is $n_{daily} = 24\ hours/r$, where r is the resolution and $h \geq r$, then $2h$ in number (h for working days and h for weekends and public holidays) error thresholds are extracted, above which a prediction identifies a burst.

As for calculating the threshold from every h-hour interval subset of error values, we extract the 99.9-th percentile of the distribution of the within-the-cluster error values. The reason why such a high percentile value is chosen is so as to account for the fact that all (known) burst events have been removed from the training and validation subsets and the validated model is fit to predict the hydraulic features of only the nominal (leak-free) conditions. Hence, anything that exceeds this threshold is something that the trained model "struggles" to understand and is therefore flagged as abnormal behaviour and ultimately identifies a burst. The distinction between a predicted "burst" and a "normal conditions" outcome is presented in Eq. (12):

$$E_t \begin{cases} \geq E_{thres,h,d} : burst \\ < E_{thres,h,d} : normal\ conditions \end{cases} \qquad (12)$$

where E_t is the value of the error function at time t and $E_{thres,h,d}$ is the error function threshold for the h-th hour interval of the day and d the relevant day type, i.e., either working day or weekend/public holiday.

Since the choice of the 99.9-th percentile of the distribution of the within-the-cluster error values is expected to have a significant effect in the sensitivity of the methodology, a posterior sensitivity analysis was performed where multiple such thresholds were investigated.

4.7 Performance assessment

To assess the performance of the methodology, both event- and value-based metrics are utilized. The event-based metrics are calculated based on the existence of non-repeated alarms, i.e., exceedances of E_{thres}, within the period defined by the repair time and one week prior to the detection of a registered leak job by the operator. Such an antecedent one week period of time is included in this definition, because water utilities do not usually become aware a burst immediately when it happens, but only after a customer has notified them or the magnitude of the water losses reflect on the entire DMA measurements. The value-based metrics are calculated on a per value basis by comparing the generated alarm instances to the leak job records, as those are defined by the start and repair time of each leak job.

A total of three event-based metrics are calculated. These are Recall$_e$, Precision$_e$ and F1-score$_e$, where subscript e denotes the event-based calculation of these metrics. Their calculation is provided in Eqs. (13)-(15).

$$Recall_e = \frac{TP}{TP + FN} \times 100\% \tag{13}$$

$$Precision_e = \frac{TP}{TP + FP} \times 100\% \tag{14}$$

$$F1 - score_e = \frac{2 \cdot (Recall_e \cdot Precision_e)}{Recall_e + Precision_e} \times 100\% \tag{15}$$

The value-based metrics calculated are Recall, Fallout and Precision. Their formulas are shown in Eqs. (16)-(18):

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{16}$$

$$Fall - out = \frac{FP}{FP + TN} \times 100\% \tag{17}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{18}$$

In Eqs. (13)-(18), TP stands for true positive (true bursts that are classified as bursts); FN stands for false negatives (true bursts that are misclassified as non-bursts), TN stands for true negative (when a burst does not occur and is not detected), and FP denotes false positive (when a burst does not occur, but it is incorrectly detected as having occurred).

Recall is useful, because it shows the proportion of data correctly identified as bursts under burst conditions. Similarly, Precision shows the proportion of the burst-flagged data that are actually registered bursts, hence relating to the confidence in the alarms. Fall-out is the probability of false alarms. Finally, F1-score is a harmonic mean of precision and recall and is a measure of composite burst detection ability best suited for imbalanced datasets. A good methodology should have high Recall, Precision and F1-score, whereas Fall-out should be minimal.

Just for the cases of detecting simulated fire hydrant leak tests, an additional metric is investigated: Detection Delay (DD). Detection Delay corresponds to the delay between the actual registered start time of a fire hydrant leak test and the first instance of an alarm, i.e., exceedance of the error threshold. This metric can actually be calculated for these two cases, because the actual start time of the leak is known, so that detection delay can be quantified. Calculation of DD is shown in Eq. (19):

$$DD = t_{detection} - t_{start} \tag{19}$$

where $t_{detection}$ corresponds to the time the burst is detected and t_{start} the time the burst actual occurred.

5. Results and Discussion

5.1 Data acquisition

The methodology of this research is tested on two types of bursts: (a) real bursts that took place on the DMAs of a UK water utility company, and (b) simulated fire hydrant leak tests that took place on a subset of these real DMAs in the UK. For the different types of bursts, the following information is available:

a) For the real DMAs, flow and pressure measurements from the inflow point of each DMA, as well as pressure measurements from the critical point at a time resolution of 15-minutes are used. In addition, leak job records with all the detected leak jobs, their detection time and date, their repair time and date and a short description of their nature is also available. The extent of both the flow and pressure measurements and the leak job records extend in the time period from October 2016 until March 2022.

b) For the simulated fire hydrant leak tests in a subset of the DMAs, the information available includes the flow and pressure measurements from the DMA inflow as before, plus the pressure measurements from the critical point of the DMA and the five to seven (depending on the DMA) additional pressure sensors throughout the DMA that were installed in early 2022 at a time resolution of 15-minutes. Furthermore, the timing of the start and end of the fire hydrant leak tests is known, as well as their discharge: $Q_{leak} = 0.8\ l/s$ for $1.5\ h$ followed by $Q_{leak} = 1.5\ l/s$ for $1\ h$, giving a total duration of $2.5\ h$.

Because each of the two types of bursts is from different operational settings, there are certain challenges that accompany their investigation. First of all, real bursts registered in the leak job records of the real DMAs most likely started before the operator detected them. As a result, it is safe to say that the real "ground truth" is very difficult to be known and the produced "leak-free" records on which the LSTM-based model is trained has inadvertently some background and/or undetected bursts present. Furthermore, in real DMAs, it is not uncommon for sensors to be recalibrated or replaced every few years, which induces additional challenges to treating the entire length of these datasets as consistent. As far as the simulated fire hydrant leak tests are concerned, these were executed in daytime and with a progressively increasing discharge, so as to not cause any unnecessary harm to the network pipes. However, such a behaviour is not always representative of real bursts.

5.2 Implementation of the methodology

After acquiring the relevant data, engineering the features and extracting the input of the model, the training-validation subsets split takes place from the engineered leak-free record, while the testing subset is extracted from the unaltered time period that follows. To create the training, validation and testing subsets, 4-day long rolling periods (including the current time step) at 30-minute resolution are extracted, creating time-series with a total length of 193 time steps (193 = 4 days × 24 hours × (60 minutes / 30 minute resolution)). Since the original data is at a resolution of 15-minutes, resampling is implemented.

The neural network is set up using the keras package in the JupyterLab Python environment using a functional model layout. The model is trained for 100 epochs using the Adam optimizer that is computationally efficient and best suited for problems with large amounts of data (Kingma & Ba, 2015). As an objective function the Mean Squared Error (MSE) is utilized, whereas callback functions for adjusting the learning rate and stopping the training, once the algorithm reaches a plateau, are also used. A callback function monitors the loss of the validation subset ensuring that when the learning rate drops below 0.0010 per 6 epochs, the learning rate will be reduced to 20% of he previously used one. Another callback function also monitors the validation loss for the purpose of stopping training, when the loss of the validation subset does not improve by at least 0.0010 within a time frame of 10 epochs.

5.3 Results in real bursts

The model is trained, validated and tested on 10 different DMAs in the UK. After training and validating the model on a leak-free portion of the dataset, the trained model is applied to the testing portion, where both leak-free periods and registered leak jobs exist. The performance of the methodology corresponding to a random run of the code script is shown in Table 1.

Table 1: Performance of LSTM-based model in real leaks

DMA	Sensors		Leak jobs	Event-based			Value-based		
	Q	P		Recall$_e$	Precision$_e$	F1-score$_e$	Recall	Fall-out	Precision
Alpha	1	2	41	29.3%	63.2%	40.0%	6.3%	0.5%	86.9%
Beta	1	2	37	29.7%	78.6%	43.1%	0.6%	0.1%	96.9%
Gamma	1	2	21	38.1%	47.1%	42.1%	9.1%	4.7%	57.1%
Delta	1	2	6	16.7%	4.2%	6.7%	16.4%	12.4%	12.2%
Epsilon	1	2	60	68.3%	65.1%	66.7%	10.6%	3.2%	98.1%
Zeta	1	2	5	40.0%	28.6%	33.3%	5.4%	0.2%	51.4%
Eta	1	2	8	62.5%	3.0%	5.7%	8.9%	3.0%	61.7%
Theta	1	2	7	57.1%	26.7%	36.4%	15.5%	2.7%	79.4%
Iota	1	2	4	50.0%	22.2%	30.8%	7.2%	0.5%	40.9%
Kappa	1	2	3	100.0%	23.1%	37.5%	6.3%	1.4%	57.2%

The methodology is applied to a total of 10 DMAs in the UK, each with periods for training, validation and testing of different length. This is caused by the requirement to have consistent flow and pressure signals that are affected the least by replacing or recalibrating the monitoring equipment.

From Table 1 it can be seen that (the event-based) Precision$_e$ is proportional to the number of leak jobs, ranging from 3.0% to 78.6%. Specifically, the correlation coefficient between Precision$_e$ and the number of leak jobs is calculated to be 0.848. A similarly high correlation coefficient of 0.750 exists between the value-based Precision and the number of leak jobs. Considering that the training, validation and testing periods of the different DMAs have length of the same order of magnitude, this phenomenon is more likely explained by the uneven level of public alertness, which is usually the identifier of pipe bursts.

This claim is supported by the land use cover of the different DMAs. Namely, DMAs Delta and Eta that exhibit the worse performance are covered by agricultural fields in more than 90% of their total area, whereas DMAs Alpha, Beta and especially Epsilon are heavily urbanized. This is a strong indication that a lot of actual pipe bursts in the rural DMAs go completely unnoticed and are not registered in the leak job records. This has a two-fold impact on our methodology. First, not all actual leak jobs are removed for training and validation, thus impairing the ability of the model to "learn" leak-free normal behavior only. Second, even if the model is trained well enough, the possible existence of multiple unregistered leaks in the testing subset leads to an overwhelming number of False Positives, which should be in fact be labeled as True Positives. This means that unreported leak jobs can account for the different overall performance across the multiple DMAs investigated here.

Another interesting observation on Table 1 values is the fact that value-based Recall is usually significantly lower than event-based Recall$_e$. This means that during periods of registered leak jobs, the model prediction error does not always exceed the calculated error threshold, but only periodically. If we examine DMA Delta, where Recall$_e$ = 16.7%, we see that only 1 of the six registered leak jobs is detected. As can be seen in Fig. (6), where the period of time corresponding to that one detected leak job is presented, the alarms, i.e., exceedances of the error threshold, are not persistent, but rather intermittent, taking place in late morning hours, after the morning peak. This is a significant limitation of the threshold-based approach. It is possible that had the tolerance reflected on the error threshold been different, the behaviour of the alarms would also have changed. Hence, sensitivity analysis is necessary.

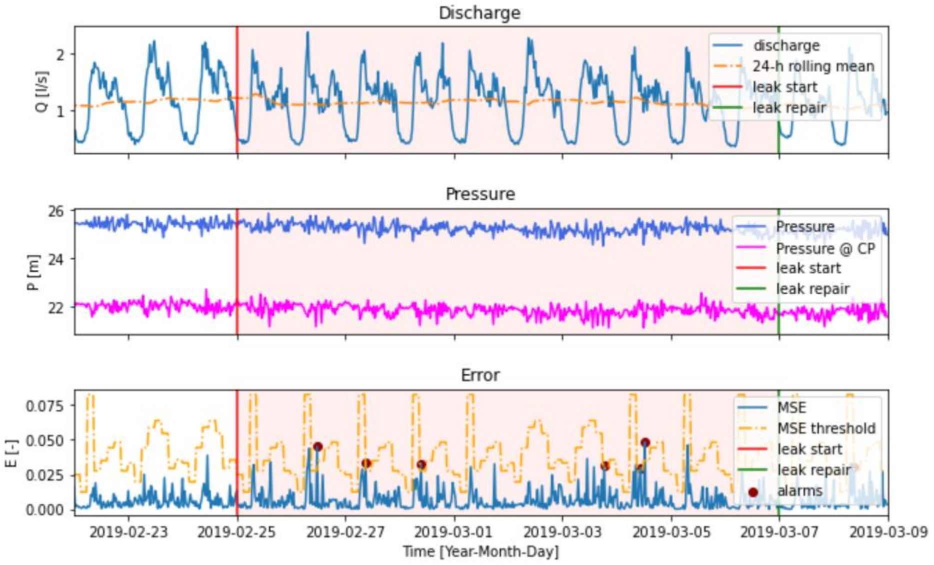

Figure 6: Investigation of the period of time around the one detected leak of DMA Delta

5.4 Sensitivity analysis

To study the effect time resolution has on the burst detection performance of our methodology, the entire process of training, validation and testing is repeated just for a subset of DMAs. It is the authors' decision to pursue this for just two DMAs and actually for those that some may render to correspond to be the best and worst performance. Based on the results of Table 1, the best performance of the model is found to be in DMA Beta, where Precision$_e$ = 78.6% and Fall-out = 0.1%. As for the worst performing DMA this is most likely Eta, that has Precision$_e$ = 3.0% and F1-score$_e$ = 5.7%. Different combinations of time resolution (15-min, 30-min and 60-min) and n-length of the input hydraulic feature time-series (1, 2, 3, …, 7 days) are investigated, while always keeping the length of the time-series less than 250 values, so that LSTM cells do not get overwhelmed by them. Tables 2 and 3 show the event- and value-based performance metrics respectively for DMAs Beta and Eta compared to the "standard" length $n = 4\ days$ and time resolution $r = 30\ min$ used across the DMAs.

Table 2: Event-based metrics of sensitivity analysis of LSTM-based model for different length of input time-series and time resolution

Event-based KPIs		DMA					
		Beta			Eta		
Recall$_e$ [%]		Time resolution [min]			Time resolution [min]		
		60	30	15	60	30	15
Length of LSTM-fed time-series [days]	7	35.1%			50.0%		
	6	37.8%			62.5%		
	5	29.7%	27.0%		75.0%	62.5%	
	4	37.8%	29.7%		50.0%	62.5%	
	3	35.1%	24.3%		50.0%	75.0%	
	2	37.8%	32.4%	37.8%	62.5%	75.0%	75.0%
	1	37.8%	29.7%	27.0%	62.5%	62.5%	75.0%
Precision$_e$ [%]		Time resolution [min]			Time resolution [min]		
		60	30	15	60	30	15
Length of LSTM-fed time-series [days]	7	76.5%			5.8%		
	6	66.7%			6.0%		
	5	61.1%	71.4%		4.0%	2.9%	
	4	63.6%	78.6%		4.4%	3.0%	
	3	65.0%	69.3%		6.6%	2.6%	
	2	73.7%	70.6%	93.3%	3.9%	1.7%	10.3%
	1	70.0%	57.9%	83.3%	3.7%	3.1%	8.8%
F1-score$_e$ [%]		Time resolution [min]			Time resolution [min]		
		60	30	15	60	30	15
Length of LSTM-fed time-series [days]	7	48.2%			10.4%		
	6	48.3%			10.9%		
	5	40.0%	39.2%		7.5%	5.5%	
	4	47.4%	43.1%		8.0%	5.7%	
	3	45.6%	36.0%		11.6%	5.0%	
	2	50.0%	44.4%	53.8%	7.3%	3.4%	18.2%
	1	49.1%	39.3%	40.8%	7.0%	5.8%	15.8%

Table 3: Value-based metrics of sensitivity analysis of LSTM-based model for different length of input time-series and time resolution

Value-based KPIs		DMA					
		Beta			Eta		
Recall [%]		Time resolution [min]			Time resolution [min]		
		60	30	15	60	30	15
n-length of input timeseries [days]	7	1.9%			5.4%		
	6	1.6%			7.6%		
	5	1.5%	0.3%		14.1%	10.8%	
	4	1.4%	0.6%		9.7%	8.9%	
	3	1.3%	0.1%		4.9%	13.5%	
	2	1.3%	0.5%	1.7%	12.4%	22.4%	18.7%
	1	1.0%	0.4%	0.2%	10.5%	8.8%	19.2%
Fall-out [%]		Time resolution [min]			Time resolution [min]		
		60	30	15	60	30	15
n-length of input timeseries [days]	7	0.1%			2.4%		
	6	0.2%			3.1%		
	5	0.2%	0.0%		7.5%	3.5%	
	4	0.2%	0.1%		3.0%	3.0%	
	3	0.2%	0.0%		2.4%	4.8%	
	2	0.1%	0.1%	0.0%	5.9%	10.7%	2.7%
	1	0.2%	0.1%	0.0%	4.7%	3.1%	2.5%
Precision [%]		Time resolution [min]			Time resolution [min]		
		60	30	15	60	30	15
n-length of input timeseries [days]	7	97.3%			55.1%		
	6	94.0%			56.6%		
	5	94.4%	92.3%		50.4%	62.6%	
	4	93.4%	96.9%		63.5%	61.7%	
	3	93.4%	84.0%		52.5%	60.2%	
	2	95.1%	94.4%	98.9%	53.1%	53.2%	65.1%
	1	93.0%	87.7%	97.3%	54.7%	60.8%	68.7%

The effect of using different time resolution and/or different length of the input time-series is not negligible. As can be seen in Tables 2 and 3, coarser resolution leads to significantly higher values of the Fall-out, which translates to less confidence on the alarms. This is most likely caused by the fact that instantaneous exceedances of the error threshold are proportionally more significant compared to the same number of instances in datasets of finer resolution.

Furthermore, in 60-min resolution, the model may not have adequate information to “decode” the short-term dynamics of bursts, because pressure and flow measurements are aggregated to 1-hour intervals. Even though there is a slight increase in the values of both event-based Recall$_e$ and value-based Recall for coarser resolution, it is maybe preferable to sacrifice the detection of a handful of bursts, for the sake of superior confidence in the alarms, i.e., higher Precision$_e$, Precision and F1-score$_e$. It can therefore be stated that a combination of 2-day long time-series at 15-min resolution is superior to the 4-day long time-series and 30-min resolution that was initially used. This is also supported by the better scores across all the performance metrics.

In addition to the sensitivity of the method to the time resolution, it is also prudent to check how the choice of the 99.9th percentile threshold of the error function affects the resulting alarms. To this end, we examine how Precision$_e$, Recall$_e$, F1-score$_e$ and the value-based Precision change with different values of this threshold. This analysis takes place for the best performing DMA, i.e., DMA Beta and its results are displayed in Figs. (7)-(10).

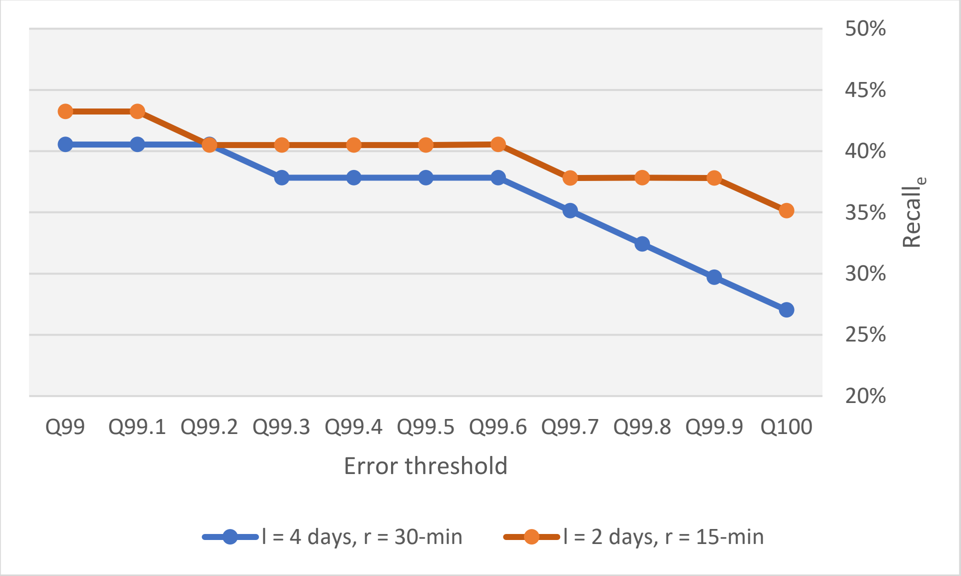

Figure 7: Recall$_e$ sensitivity to the error threshold

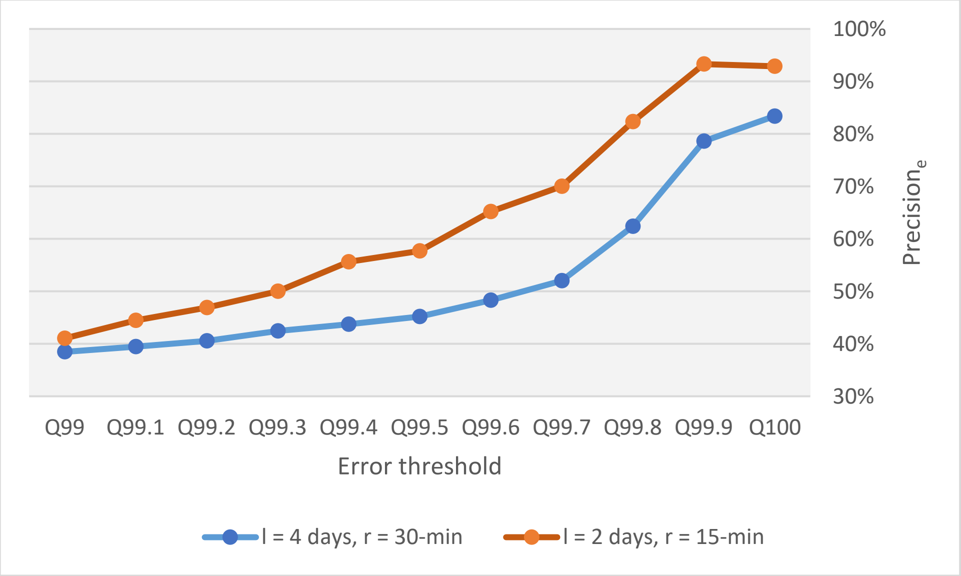

Figure 8: Precision$_e$ sensitivity to the error threshold

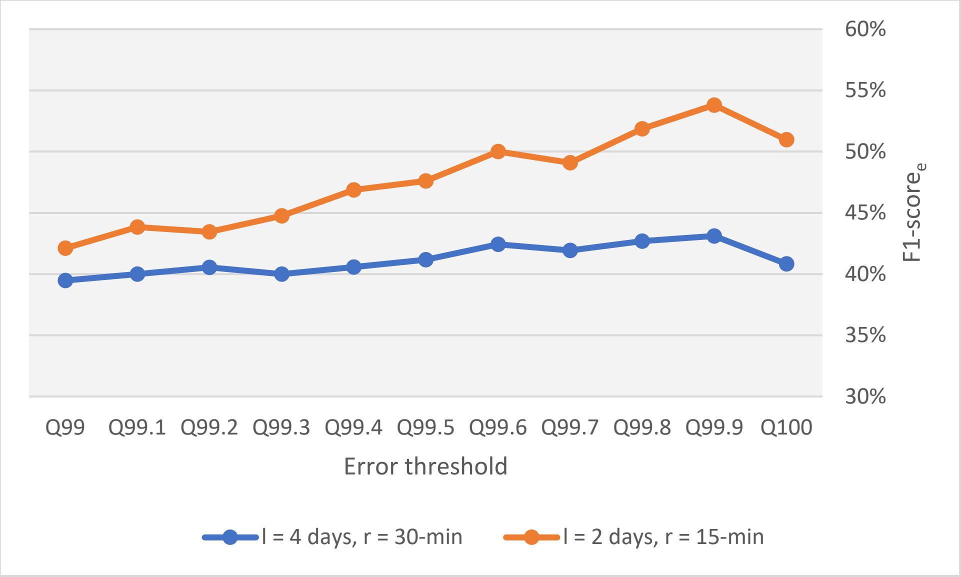

Figure 9: F1-score$_e$ sensitivity to the error threshold

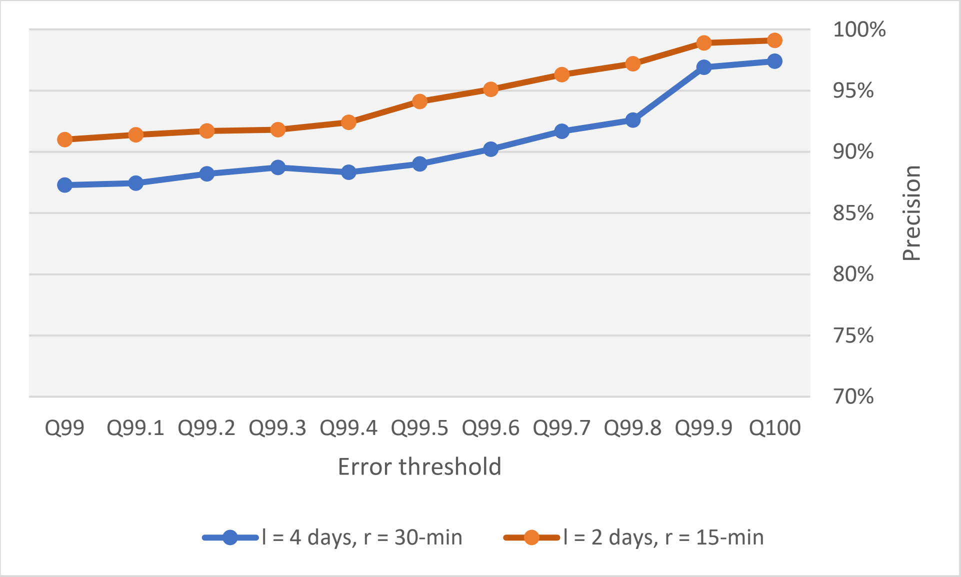

Figure 10: Precision sensitivity to the error threshold

Figs. (7)-(10) show the effect different percentile thresholds of the validation error have on the burst detection performance. As was expected, lower thresholds lead to more leak jobs being detected and an overall higher Recall$_e$. However, this is coupled by additional false alarms that severely impact the Precision, which plummets from higher than 80% to lower than 50%. This trade-off is better quantified though the composite F1-score$_e$, which acquires its greatest value for the 99.9th percentile for both combinations of n-day long time-series and resolutions. As for the value-based precision, lowering the error threshold leads to lower values of the metric, however not to the extent experienced with the same event-based metric.

5.5 Results in fire hydrant leak tests

In March 2022 the UK water utility company performed fire hydrant leak tests in selected DMAs to assess the performance of the developed burst detection methodology. These leak tests were carried out in a controlled manner, with a fixed duration and discharge (see Section 5.1). The burst discharge relative to the mean DMA inflow q_{burst} is shown in Table 4.

Table 4: Artificial fire hydrant leak size compared to mean DMA inflow

DMA	Beta		Delta		Zeta	
$Q_{DMA,mean}$	7.2 l/s		1.3 l/s		2.5 l/s	
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
q_{burst}	11%	21%	62%	115%	32%	60%

In early 2021 new pressure sensors were installed in the DMAs where the fire hydrant leak tests took place. Hence the training, validation and testing subsets of the model are limited to the period after that, so as to preserve a consistent behaviour of the system, on which the model is to be trained. In DMAs Beta and Delta, the fire hydrant leak tests were performed at 10 March 2022, whereas in Zeta, it was performed in 15 March 2022. Furthermore, since results from the sensitivity analysis showed that 2-day long time-series and a resolution of 15-minutes provides the best results, this is utilized in the study of the fire hydrant leak tests.

Three different scenarios are initially investigated for the assessment of the performance on detecting these artificial fire hydrant leak tests. Scenario A, where only the 3 originally installed sensors (1 for flow and 2 for pressure) are used. Scenario B, where all the available sensors are utilized, but with a severely limited training and validation periods, due to the very recent installation of the additional pressure sensors. Scenario C, where all the available sensors are utilized, but with utilizing transfer learning of the model from scenario A, and just fine-tuning the expanded model for a period of 1 month.

Table 5: Performance on fire hydrant leak detection for scenarios A, B and C

Scenario A: Additional sensors not utilized								
DMA	Sensors		Leak tests		Value-based			
	Q	P	Performed	Detected	Recall	Fall-out	Precision	DD
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-min
Scenario B: Additional sensors utilized without transfer-learning								
DMA	Sensors		Leak tests		Value-based			
	Q	P	Performed	Detected	Recall	Fall-out	Precision	DD
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-min
Zeta	1	2+5	1	1	7.4%	2.2%	6.0%	60-min
Scenario C: Additional sensors utilized with transfer-learning								
DMA	Sensors		Leak tests		Value-based			
	Q	P	Performed	Detected	Recall	Fall-out	Precision	DD
Beta	1	2+7	1	1	11.1%	2.7%	1.9%	15-min
Delta	1	2+5	1	1	90.9%	22.6%	1.9%	30-min
Zeta	1	2+5	1	1	36.4%	2.8%	14.3%	30-min

Table 5 shows that for scenario A, i.e., when the additionally installed pressure sensors are not utilized, only 1 of the 3 performed fire hydrant leak tests is detected and that only after a delay of 60-min. This bad behavior is explained by the fact that the leak tests took place far away from both the inflow and critical points where sensors are installed, as well as the fact that there is significant fragmentation of the training and validation subsets, due to “benching” of the pressure signals caused by operator-induced alterations, as well a sensor replacements/recalibration.

To account for the increased spatial coverage offered by the additionally installed pressure sensors, scenario B utilizes all available sensors. However, in this case, only 2 out of the 3 leak tests are detected and that with a delay of 30- and 60-min. This is caused by the lack of proper training, which is the result of the very young age of the additionally installed pressure sensors. Hence, the model gets fitted to not only a small dataset, but also a very limited one with respect to (inter-)annual seasonality. Had the testing subset been expanded beyond the month of March, the performance would sure be much worse.

Scenario C makes use of the inherent flexibility of the LSTM neurons, that enable the seamless transfer of knowledge, I.e., weights, between pre-trained nodal connections and newly created ones. Hence, in Scenario C it is the previously trained weights that are linked to the pressure sensor installed at the inflow point that enable the weights of all the additionally installed sensors to have initial values much closer to the optimal ones, than the default ones that are 0. In Scenario C all three fire hydrant leak tests are detected, the one at DMA Beta with a delay of just 15-min, i.e., the very first time step, and at DMAs Delta and Zeta with a delay of 30-min, i.e., the second time step of the simulated leak.

However, what is striking from the performance metrics presented in Table 5 is the excessively high values of the Fall-out for DMA Delta. Upon closer investigation of the raw time-series of all the available sensors in that area it was found that the pressure sensor installed at the critical point exhibits irrational entries just prior to the leak test with its values increasing by more than 30 m. Had these measurements been real, that would have been reflected in the recordings of the other pressure sensors as well.

To tackle this problem of faulty pressure entries and thanks to the natural scalability of the LSTM model, two additional scenarios are investigated. Scenarios D and E are identical to scenarios B and C respectively, with the exception that the nodal connections corresponding to the problematic pressure sensor at the critical point are removed. The results are shown in Table 6.

Table 6: Performance on fire hydrant leak detection for scenarios D and E

Scenario D: Additional sensors utilized without transfer-learning and P_{CP} removed								
DMA	Sensors		Leak tests		Value-based			
	Q	P	Performed	Detected	Recall	Fall-out	Precision	DD
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-min
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								
DMA	Sensors		Leak tests		Value-based			
	Q	P	Performed	Detected	Recall	Fall-out	Precision	DD
Beta	1	1+7	1	1	9.1%	3.3%	1.3%	15-min
Delta	1	1+5	1	1	27.3%	1.7%	7.0%	30-min
Zeta	1	1+5	1	1	18.2%	4.4%	1.9%	60-min

As was expected, the results on the performance of the methodology on DMA Delta, where the problematic pressure sensor was installed, are better for both scenarios, since the Fall-out plummets and Precision in increased. For the same DMA there does not seem to be an improvement in terms of detection delay and in fact the value-based Recall is reduced. That is however not a serious drawback, since the confidence to the alarms moves in the opposite direction to the Fall-out.

Another interesting finding is the fact that for DMAs Beta and Zeta that did not exhibit any false pressure measurements, the exclusion of the pressure sensor at the critical point deteriorates the performance, with lower Recall and Precision and higher Fall-out. Especially for the case of DMA Zeta in scenario D, the exclusion of this one pressure sensor is pivotal and leads to the fire hydrant leak test going completely unnoticed. This signifies the importance of the spatial coverage of every single sensor and the fact that unnecessary removal of information streams has negative influence in the overall performance of burst detection.

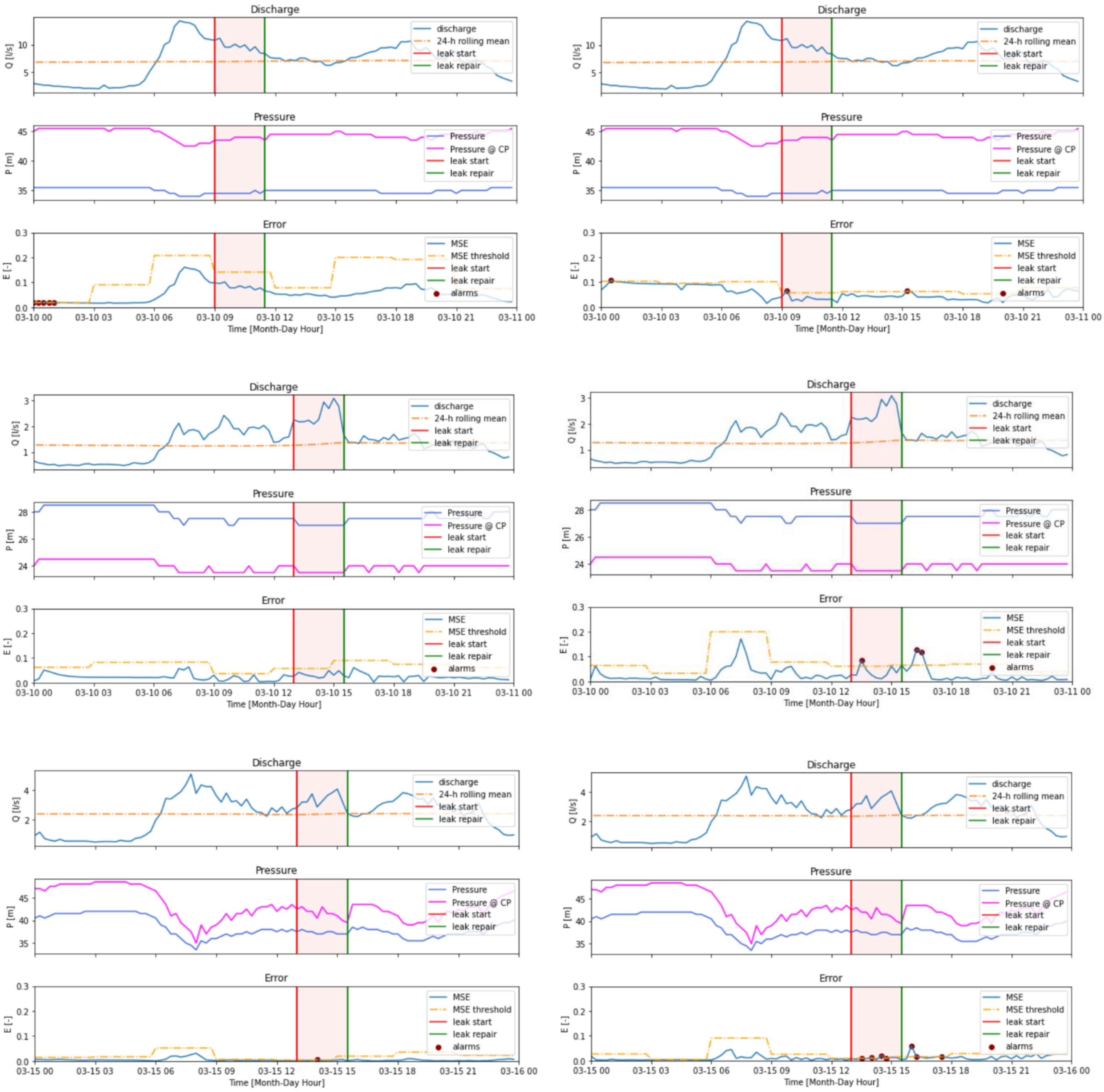

Figure 11: Fire hydrant leak tests in DMAs Beta (top), Delta (middle) and Zeta (bottom).. Left sub-figures show the result of running the model with two pressure signals as input in scenario A. Right sub-figures show the result of running the model with all the pressure sensors available as input after utilizing transfer learning in scenario C. Sub-figures in top row show the discharge at the inflow of the DMA. The middle row sub-figures show the pressure signals at the inflow of the DMA and the critical point. The lower sub-figures show the MSE (error) function, the variable error threshold, the start and repair time of the fire hydrant leak tests and the raised alarms.

Fig. (11) presents a snapshot of the fire hydrant leak tests in the three selected DMAs and the whole day in which they took place. An interesting finding from a visual inspection of these figures is the fact that there appear to be residual alarms, i.e., exceedances of the error threshold, after the leak is repaired (stopped in this case) and the system returns to normality. This is most likely a sensitivity of the LSTM network and it is something of which the operator should be aware. Nevertheless, for the purposes of this study, such posterior alarms are flagged as false and are reflected in Fall-out.

5.6 Comparison to other works

The methodology applied in this study is closer in scope and neural networks used to the one of (Wang et al., 2020), who used a purely LSTM model to detect both simulated and synthetic bursts in a single real-life DMA. In the simulated leak tests, (Wang et al., 2020) showed that their LSTM model that used data of 5-min resolution, was able to detect them in two time steps, i.e., 10-min time. This is comparable to our findings (see Table 5). However, in one of the three fire hydrant leak tests that we apply our model, the leak-test are identified instantaneously, i.e., in the very first time-step. The synthetic leaks examined by Wang et. al. (2020) are not really comparable to our findings on real leaks, since their behaviour is completely different.

Lee & Yoo (2021) implemented a less similar-to ours LSTM methodology with only flow data to detect a single leak on a part of a water distribution system feeding a city. Utilizing a total of 6 flow sensors, Lee & Yoo (2021) were able to detect the one leak with a sensor-based accuracy ranging from 61.53% up to 99.79% and a false positive rate ranging from 0.11% to 29.88%. Compared to our methodology, we believe that this inferior performance may be attributed to the exclusion of pressure and time features, the small training and validation subsets of only few-days length, and finally the fact this was only tested on a single leak.

Comparing the performance of burst detection on the same dataset of real leaks is also important. That is why methodologies outside the boundaries of RNNs need to be assessed. To do this, it useful to comprehend that burst detection in water distribution systems can be basically “stripped down” to anomaly detection. That it because bursts are inconsistencies, i.e., anomalies, in an otherwise normal-behaving time-series. Taormina & Galelli (2018) were the first to propose the use of Autoencoders (AE), a novel deep learning neural network approach, to detect cyber-attacks in water time-series. AE is composed of interlinked encoder and decoder and is aimed at comprehending a compressed representation of high-dimensional input data by means of reconstructing it through an error-minimization approach. 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, which is what makes it state-of-the-art.

Hyperparameter tuning is very important for the AE as it is with LSTM networks. To this end, 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. To this end, an Autoencoder with successive layers of 64, 32, 16, 32 and 64 neurons each resulted to have the optimal structure for this problem.

All the hydraulic and domain (time) features are used as both input and output. Specifically, the AE is fed 4-day long time-series of the flow and pressure signals from the DMA inflow, pressure signal from the critical point and the two time features. Due to the self0supervised nature of the AE, the goal of training it is recreating the input as output with the highest fidelity possible. The Adam optimizer is used again, as is the case with the LSTM-based model, and the ReLu function is employed to activate the neurons. The results of applying this AE on the real leaks are provided in Table 7.

Table 7: Performance of the Autoencoder in real leaks

DMA	Sensors		Leak jobs	Event-based			Value-based		
	Q	P		Recall$_e$	Precision$_e$	F1-score$_e$	Recall	Fall-out	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%

Based on the results of Table 7, it is seen that 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 1). All the metrics, both event- and value-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 (which is better) values of Fall-out for DMAs Delta, Epsilon, Eta and Iota. However, the fact that they 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. Hence, the LSTM-based model seems to exhibit an overall better performance in both real leaks and simulated fire hydrant leak tests compared to existing state-of-the-art methodologies.

6. Conclusions

This work contributes a neural network-based detection algorithm for identifying pipe bursts at the DMA scale of water distribution systems. The main idea on which the methodology relies is the development of a data-based model that is trained to predict the patterns of all the hydraulic features (namely flow and pressure) during normal operating conditions and as such reports high prediction errors when fed data that correspond to pipe bursts having taken place. The implementation of this methodology is made possible by the use of an LSTM-based model, a special kind of recurrent neural network that is characterized by its feedback connections.

Using LSTM neurons in our methodology equips the detection algorithm with two major advantages: it offers superior predictive power by incorporating the history of a significant length of past information without exhibiting the exploding/vanishing gradient problem upon training and it assists with adding/removing information streams seamlessly according to the availability/suitability of the installed monitoring sensors, since this flexibility is embedded in the LSTM neurons. Furthermore, the inherent flexibility of LSTM cells enables utilizing transfer learning under conditions of data scarcity, when a full-scale training is not possible.

The current LSTM approach is not the only LSTM-based methodology in the field of burst detection. Other researchers have previously worked on this (see Section 2) getting very promising results. However, this is the first LSTM-based methodology to be tested on both real leaks and simulated fire hydrant leak tests. Furthermore, the sensitivity analysis performed with respect to the temporal resolution of the data fed and the incorporation of additional information streams from sensors installed at a later time by implementing transfer learning and fine-tuning are also points of novelty of this research.

Results obtained from testing the algorithm in real leak job records showed that the developed data-driven algorithm works reasonably well. Performance inconsistencies however among different DMAs showed that utilizing an accurate leak job record is of paramount importance. That is for training the model on leak-free periods, as well as for correlating alarms to actual pipe bursts in the testing phase. Limited public awareness and rural DMAs render a lot of bursts to go completely unnoticed, with a profound impact on the performance of this and other data-driven approaches. Sensitivity analysis showed that finer data resolution leads to better overall performance with event-based Precision_e reaching 93.3%, as the abrupt disturbances caused by pipe bursts are not averaged out in coarser measurements. This is possible by the inclusion of domain (time) features in a different "stream" of input, separate to the hydraulic features, that enables the model to better understand the diurnal and weekly variations, as well as the expected abnormalities caused by bank holidays. Furthermore, the use of a variable error threshold mirroring the variable water consumption patterns expected during the day enables a varying degree of sensitivity tailored to each DMA and weekday. This further increases the robustness of this burst detection approach and makes it applicable to real life settings.

Results acquired from testing the algorithm in the simulated fire hydrant leak tests reveal how important the location of the leak in relation to the sensors is to its detection. It is proven that detecting the tests in time, or even at all, may be hard if theses that take place far away from installed sensors. However, by incorporating information from additionally installed sensors scattered throughout the DMA, all leak tests with a flow rate down to 11% of the mean DMA inflow, are detected within one or two time steps. The mere inclusion of additional information streams is possible thanks to the embedded flexibility of the LSTM cells, in combination with transfer learning prior information corresponding to an existing pressure sensor and fine-tuning the model.

Limitations of the current approach include the requirement for training the model on leak-free datasets, which in turn presupposes the existence of accurate leak job records that include all the bursts taking place. The latter is rather utopic to strive for, especially in rural sparsely populated areas and/or DMAs with a limited number of monitoring sensors in relation to their overall extent. In addition, the methodology is sensitive to sensor recalibration/replacement and special care needs to be taken to identify periods of consistent measurements on which the model is been applied. As was discussed in the application of the algorithm in areas with problematic sensor measurements, it is possible to exclude them, but only if one is certain of this. Otherwise, such an action will have adverse effects on the spatial coverage of the monitored area and consequently to the burst detection performance.

Naturally, a thorough investigation of the universal applicability of the proposed methodology to a variety of real-life water distribution systems necessitates further exploration. For instance, it is required to investigate what is the impact of the typical topology of the installed sensors (at the inflow and the critical points) found in the DMAs in focus in the UK. In terms of calculating the prediction error, it would also be interesting to assess what the impact of the use of a quadratic function is and if the use of another function will be beneficial. Furthermore, modern monitoring equipment makes possible flow and pressure measurements at 1-min interval. It is interesting, if not necessary, to assess how the performance of burst detection changes with such high resolution data. In addition, meteorological extremes have shown to greatly affect water consumption, especially in areas with extensive vegetative cover, such as urban gardens. Assessing the impact of the algorithm to temperature and even incorporating it in future approaches may be promising. Finally, this method implicitly assumes that water consumption behavior is stationary and only affected by pipe failures and repair jobs. Although this may be true for short time horizons, urbanization, quality of life improvements and climatic variability may render this moot for long-term horizons. Hence, it may be interesting to assess the impact such slowly developing phenomena have on the current burst detection approach.

Code Availability

The Python code scripts created for this study can be accessed in: https://github.com/konglynis/thesis_repo.git.

Acknowledgments

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 Haochen Zhang from Royal HaskoningDHV for his advice and insight into the operation of water distribution systems and machine learning.

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.

Bakker, M., Trietsch, E. A., Vreeburg, J. H., & Rietveld, L. C. (2014). Analysis of historic bursts and burst detection in water supply areas of different size. *Water Science and Technology: Water Supply, 14*(6), 1035-1044. doi:doi.org/10.2166/ws.2014.063

Bakker, M., Vreeburg, J., Rietveld, L., & Van der Roer, M. (2012). Reducing customer minutes lost by anomaly detection? *14th Water Distribution Systems Analysis Conference.* Adelaide, Australia.

Bjerke, M. (2019). *leak detection in water distribution networks using gated recurrent neural networks (Master's thesis).* Trondheim, Norway: Norwegian University of Science and Technology.

Brdys, M., & Ulanicki, B. (1996). Operational control of water systems: Structures, algorithms and applications. *AutomaticaSource:Elsevier ScienceDirect Journals, 32*(11), 1619.

Caputo, A. C., & Pelagagge, P. M. (2003). Using neural networks to monitor piping systems. *Process Safety Progress, 22*(2), 119-127.

Casillas Ponce, M. V., Garza Castanon, L. E., & Cayuela, V. P. (2014). Model-based leak detection and location in water distribution networks considering an extended-horizon analysis of pressure sensitivities. *Journal of Hydroinformatics, 16*(3), 649-670.

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. *Journal of Hydroinformatics, 23*(3), 453-465.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint*, 1409.1259. doi:https://doi.org/10.48550/arXiv.1409.1259

Farley, M., Water, S., Supply, W., Council, S. C., & WHO, W. H. (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.

Fausett, L. (1994). *Fundamentals of neural networks.* Englewood Cliffs, NJ, 7632: Prentice Hall Intenational.

Fox, S., Shepherd, W., Collins, R., & Boxall, J. (2016). Experimental quantification of contaminant ingress into a buried leaking pipe during transient events. *Journal of Hydraulic Engineering, 142*(1). doi:10.1061/(ASCE)HY.1943-7900.0001040

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural computation, 9*(8), 1735-1780.

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.

Hutton, C., & Kapelan, Z. (2015). A probabilistic methodology for quantifying, diagnosing and reducing model structural and predictive errors in short term water demand forecasting. *Environmental Modelling & Software, 66*, 87-97.

Kang, D., & Lansey, K. (2011). Demand and roughness estimation in water distribution systems. *Journal of Water Resources Planning and Management, 137*(1), 20-30.

Keras. (2020, May 12). *Developer guides: Complete guide to transfer learning & fine-tuning in Keras*. Retrieved from Keras.io: https://keras.io/guides/transfer_learning/

Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. arXiv preprint. *International Conference on Learning Representations.* San Diego, California, United States: ICLR 2015.

Lai, G., Chang, W. C., Yang, Y., & Liu, H. (.-a.-t.-1. (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. doi: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.

Morrison, J., Tooms, S., & Rogers, D. (2007). *District Metered Areas Guidance Notes.* International Water Association.

Mounce, S. R., & Boxall, J. B. (2010). Implementation of an on-line artificial intelligence district meter area flow meter data analysis system for abnormality detection: a case study. *Water Supply, 10(3), 437-444., 10*(3), 437-444. doi:doi.org/10.2166/ws.2010.697

Mounce, S. R., & Machell, J. (2006). Burst detection using hydraulic data from water distribution systems with artificial neural networks. *Urban Water Journal, 3*(1), 21-31.

Mounce, S., Day, A., Wood, A., Khan, A., Widdop, P., & Machell, J. (2002). A neural network approach to burst detection. *Water science and technology, 45*(4-5), 237-246.

Mounce, S., Khan, A., Wood, A., Day, A. W., & Machell, J. (2003). Sensor-fusion of hydraulic data for burst detection and location in a treated water distribution system. *Information Fusion, 4*(3), 217-229.

Mounce, S., Mounce, R., Jackson, T., Austin, J., & Boxall, J. (2014). Pattern matching and associative artificial neural networks for water distribution system time series data analysis. *Journal of Hydroinformatics, 16*(3), 617-632.

Oliker, N., & Ostfeld, A. (2014). A coupled classification–evolutionary optimization model for contamination event detection in water distribution systems. *Water research, 51*, 234-245.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *International conference on machine learning* (pp. 1310-1318). PMLR. Retrieved from https://proceedings.mlr.press/v28/pascanu13.html

Perelman, L., Allen, M., Preis, A., Iqbal, M., & Whittle, A. (2015). Flexible reconfiguration of existing urban water infrastructure systems. *Environmental Science & Technology, 49*(22), 13378-13384.

Pérez, R., Cugueró, M., Cugueró, J., & Sanz, G. (2014). Accuracy assessment of leak localisation method depending on available measurements. *Procedia Engineering, 70*, 1304-1313.

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 Engineering Practice, 19*(10), 1157-1167.

Romano, M., Kapelan, Z., & Savić, D. (2011). Burst detection and location in water distribution systems. *World Environmental and Water Resources Congress 2011: Bearing Knowledge for Sustainability* (pp. 1-10). Palm Springs, California: American Society of Civil Engineers.

Romano, M., Kapelan, Z., & Savić, D. (2014). Automated detection of pipe bursts and other events in water distribution systems. *Journal of Water Resources Planning and Management, 140*(4), 457-467.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85-117.

Siami-Namini, S., Tavakoli, N., & Namin, A. (2018). A comparison of ARIMA and LSTM in forecasting time series. In 2018 17th IEEE international conference on machine learning and applications (ICMLA) (pp. 1394-1401). IEEE. *2018 17th IEEE international conference on machine learning and applications (ICMLA)* (pp. 1394-1401). Orlando, Florida: IEEE.

Sophocleous, S., Savić, D., & Kapelan, Z. (2019). Leak localization in a real water distribution network based on search-space reduction. *Journal of Water Resources Planning and Management, 145*(7), 04019024. doi:10.1061/(ASCE)WR.1943-5452.0001079

Sun, C., Parellada, B., Puig, V., & Cembrano, G. (2020). Leak localization in water distribution networks using pressure and data-driven classifier approach. *Water, 12*(1), 54. doi:10.3390/w12010054

Svozil, D., Kvasnicka, V., & Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems, 39*(1), 43-62.

Taormina, R., & Galelli, S. (2018). Deep-learning approach to the detection and localization of cyber-physical attacks on water distribution systems. *Journal of Water Resources Planning and Management, 144*(10), 04018065.

Wang, X., Guo, G., Liu, S., Wu, Y., Xu, X., & Smith, K. (2020). Burst detection in district metering areas using deep learning method. *Journal of Water Resources Planning and Management, 146*(6), 04020031.

Wu, Y., & Liu, S. (2017). A review of data-driven approaches for burst detection in water distribution systems. *Urban Water Journal, 14*(9), 972-983.

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.

Ye, G., & Fenner, R. A. (2014). Weighted least squares with expectation-maximization algorithm for burst detection in UK water distribution systems. *Journal of Water Resources Planning and Management, 140*(4), 417-424.

Yu, R., Zheng, S., Anandkumar, A., & Yue, Y. (2017). Long-term forecasting using tensor-train rnns. *Arxiv*. doi:doi.org/10.48550/arXiv.1711.00073