Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 58-4 (2024) 79–84

Leak localization in an urban water distribution network using a LSTM deep neural network

Gómez-Coronel, L. [*,**] **Santos-Ruiz, I.** [**] **Blesa, J.** [*,***] **Puig, V.** [*,***] **López-Estrada, F.R.** [**]

[*] *Institut de Robòtica i Informàtica Industrial (IRI), UPC-CSIC, Llorens i Artigas, 4-6, 08028 Barcelona, Spain*

[**] *Tecnológico Nacional de México, I.T. Tuxtla Gutiérrez, TURIX-Dynamics Diagnosis and Control Group. Carretera Panamericana S/N, 29050 Tuxtla Gutiérrez (Mexico)*

[***] *Supervision, Safety and Automatic Control Research Center (CS2AC) of the Universitat Politècnica de Catalunya, Rambla Sant Nebridi 22, 08222 Terrassa, Spain*

Abstract: Given that water distribution networks are complex systems exposed to factors that induce leaks, it is necessary to implement techniques that allow to locate water leakages as accurately as possible minimizing the required instrumentation. In this paper we propose a leak localization technique based on the use of a long short-term memory (LSTM) deep neural network for classification trained with all possible leak scenarios in the network. As a case study, a real-world district metered area (DMA) is selected. The DMA is first sectorized considering the topological proximity of the nodes. Then, a LSTM is trained with pressure and flow rate data from all the possible leak scenarios in the system obtained from a hydraulic simulator model of the network. To replicate realistic measurements, uncertainty in the demand pattern, nominal water consumption and in the sensor readings is considered. Classification results are presented both for the validation during the training of the LSTM and for measured data of a real induced leak in the system.

Keywords: Leak localization, neural network, LSTM, deep learning, urban water management.

1. INTRODUCTION

Water is a very valuable natural resource. Its correct management and transportation is an important issue both for governments and water agencies around the globe. Water is mainly transported to the users using pipeline networks. Because of the size and complexity of these systems, they are exposed to different factors that induce leak appearance such as the aging and corrosion of the pipeline, which can be minimized through maintenance operations. Other factors, as stated by Carnero and Gómez (2018), such as the stealing of the fluid through illegal tappings or the water-hammer effect are more difficult or even impossible to predict. In Santos-Ruiz et al. (2018), it was mentioned that 21% of potable water worldwide is lost due to leaks, whereas in Mexico this percentage grows to 40%. According to a study conducted by Romano and Akhmouch (2019) and OECD (2016) about the cities with the greatest water leakages, the top four cities are located in Mexico. Tuxtla Gutierrez leads this top with a loss percentage close to 70%. According to Salehi (2022), currently the 25% of big cities are experiencing some sort of hydric stress and it is expected that the percentage of global water consumption will grow to 55% in coming years. The aforementioned facts puts into relevance studies related to water leakage detection and localisation.

1.1 Problem statement

Two tasks are of interest in the study of leakages of water distribution systems: 1) the localization tasks whose objective is to accurately provide with the location of the node water is leaking and 2) the diagnosis task, where additionally to the location of the leaks their magnitude is also to be determined. Ideally, leaks should be located as promptly as possible to perform the corresponding corrective actions. However, according to Santos-Ruiz et al. (2023), this task is non trivial due to the elevated number of possible positions where the leaks can be located or even if it is taken into consideration that the greater part of pipeline systems are buried underground. This makes difficult the visual identification of leaks and justfies the necessity of implementing methods able to localise leaks relying on the minimum number of sensors as possible.

1.2 Related works

Several different methods have been proposed in the literature both for leak localization and diagnosis. On the one hand, there are methods that rely on the use of specialized hardware. For example Martini et al. (2017) proposed the use of optic fiber, whereas Huang et al. (2007) demonstrated an application using acoustic reflectometry.

10.1016/j.ifacol.2024.07.197

However, these methods involve an elevated cost factor and thus are not the optimal choice for all applications. On the other hand, methods have been proposed using various algorithms based on measurements obtained from the operation of the system. Particularly, it was noted by Fu et al. (2022) that methods based on the use of deep learning (particularly different artificial neural networks architectures) have been gaining importance in recent years. Mainly two focuses can be identified from the literature review: classification-based methods and prediction-based methods. The main difference between the two approaches is that in the prediction scheme the objective is to forecast the evolution of the states of the system to stablish wether or not a faulting condition is found whereas in the classification-based scheme learning models are trained with labeled data from normal and abnormal conditions to distinguish different leak scenarios. The main drawback is that collecting and labeling such amounts of data is often difficult. Thus, hydraulic simulation software is regularly used to create synthetic datasets. This approach is seen in works such as Javadiha et al. (2019) in the training of neural networks and Zhou et al. (2019) for application with self-encoders. The authors in Javadiha et al. (2019) propose a convolutional neural network (CNN) to classify leak events in a simplified hydraulic model of the Hanoi district metered area (DMA). The CNN was trained with pressure maps of all possible leak locations, obtaining a maximum accuracy of 94%. An enhanced version of this work was presented by Romero et al. (2020) in which measurements from eight pressure head meters are transformed into 8×8 figures used to train a CNN with information of all possible leak scenarios in a test network. Each possible leak position has an associated signature, where nodes close to each other are clustered together given that they present similar signatures. Images are converted using a technique based on the Gramian Angular Field (GAF). Even if satisfactory results were found, a significant degradation of the accuracy of the method was observed once uncertainty in the measurements is included. A more recent implementation was presented by Romero et al. (2022) where uncertainty of $\pm 5\%$ was considered on user demands, the roughness coefficient of the pipeline and also assumed to be present in sensor readings. It was reported that even if results outperform what was previously presented in the state of the art, work was still pending in the analysis of how uncertainty affects the accuracy of the method.

Overall, it was noted by Fu et al. (2022) during the review of the state-of-the-art that CNNs are the predominant deep learning algorithm in the task of leak detection and localization. Thus, further attention should be focused in other deep learning architectures such as long short-term memory (LSTM) to explore the spatial and temporal relation between the variables. This paper shows an implementation of a LSTM deep neural network (DNN) trained with synthethic data for leak localization in a real-world water distribution network. The trained network was tested both with simulated and experimental data and satisfactory results were found. This paper is structured as follows: Section 2 introduces theoretical background regarding the LSTM network architecture. Section 3 presents the description of the case study and explains the proposed LSTM architecture. Section 4 presents with both simulation and experimental results. Finally, some conclusions and future work proposals are mentioned in Section 5.

2. THEORETICAL BACKGROUND

Artificial Neural Networks (ANNs) are popular machine-learning algorithms which emalute the learning mechanism of living beings. The biological mechanism of neurons in the brain is simulated through these computation units Aggarwal (2018). As stated by Wang et al. (2021) and Kong et al. (2018), a kind of ANN architectures widely used for capturing dynamic behaviors in data time-series are recurrent neural networks (RNNs). The mathematical expression a RNN is given by Wang et al. (2021) as:

$$\mathbf{h}^t = \sigma\left(\mathbf{W}\mathbf{x}^t + \mathbf{U}\mathbf{h}^{t-1} + \mathbf{b}_h\right) \tag{1}$$

$$\mathbf{y}^t = \sigma\left(\mathbf{V}\mathbf{h}^t + \mathbf{b}_o\right) \tag{2}$$

Here, $\mathbf{x}^t$ and $\mathbf{y}^t$ are the input and output vectors, respectively. $\mathbf{h}^{t-1}$ and $\mathbf{h}^t$ are the hidden states at times $t-1$ and t. Both the weight matrices $\mathbf{W}$, $\mathbf{U}$ and $\mathbf{V}$ and the bias vectors $\mathbf{b}_h$ and $\mathbf{b}_o$ are parameters to be adjusted during the training of the network. Finally, σ is the sigmoid activation function.

An LSTM network is an architecture of RNN that, as stated by Sangiorgio et al. (2021), introduces three additional gates: 1) A "forget gate" to control whether to keep the content in previous memory space, 2) an input gate to control whether the input vector is written into the internal memory space and 3) an output gate to control whether the value in the memory space should be outputted. Thus, according to Wang et al. (2021), the mathematical expression of a LSTM block is

$$\mathbf{f}^t = \sigma\left(\mathbf{W}_f\mathbf{x}^t + \mathbf{U}_f\mathbf{h}^{t-1} + \mathbf{b}_f\right), \tag{3}$$

$$\mathbf{i}^t = \sigma\left(\mathbf{W}_i\mathbf{x}^t + \mathbf{U}_i\mathbf{h}^{t-1} + \mathbf{b}_i\right), \tag{4}$$

$$\mathbf{m}^t = \tanh\left(\mathbf{W}_m\mathbf{x}^t + \mathbf{U}_m\mathbf{h}^{t-1} + \mathbf{b}_m\right), \tag{5}$$

$$\mathbf{o}^t = \sigma\left(\mathbf{W}_o\mathbf{x}^t + \mathbf{U}_o\mathbf{h}^{t-1} + \mathbf{b}_o\right), \tag{6}$$

where $\mathbf{f}^t$, $\mathbf{i}^t$ and $\mathbf{o}^t$ respectively represent the output vectors for the forget, input and output gates, $\mathbf{m}^t$ is the adjustment at the input, $\mathbf{W}_*$ and $\mathbf{U}_*$ are weight matrices with corresponding $\mathbf{b}_*$ bias vectors, where the subscript $*$ is either f, i, m or o. Finally, the memory space $\mathbf{c}^t$ and hidden state $\mathbf{h}^t$ are computed as $\mathbf{c}^t = \mathbf{i}^t \circ \mathbf{m}^t + \mathbf{f}^t \circ \mathbf{c}^{t-1}$ and $\mathbf{y}^t = \mathbf{h}^t = \mathbf{o}^t \circ \tanh\left(\mathbf{c}^t\right)$. The architecture of an LSTM block is presented in Figure 1.

3. METHODOLOGY

3.1 Description of the case study

The case study here proposed is a DMA of the water distribution network of Madrid, Spain (see Figure 2).

This system is composed of 312 connection nodes (cyan colored) and one reservoir node that feeds with water the system. The pipeline system is approximately 14 km in length and the diameter of the pipes vary between the 80 mm to 350 mm. Pressure head meters are installed at 10 different nodes in the system (highlighted yellow in Figure 2a). Outflow demanded from the reservoir is also monitored. A fire hidrant (highlighted red in Figure 2a) is used to simulate a leak event with a magnitude $Q_{\text{leak}} \approx$

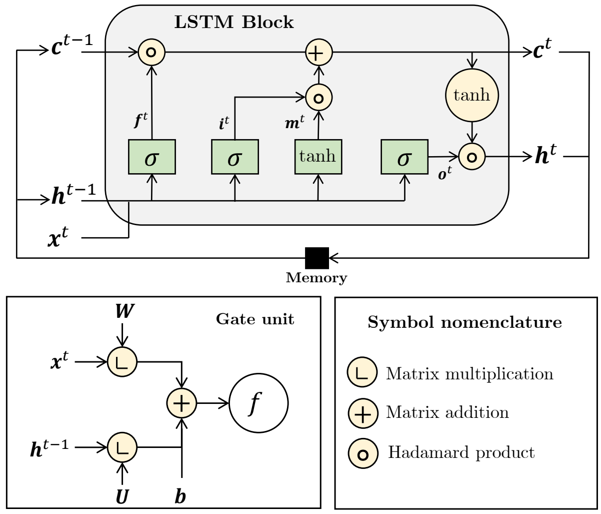


Fig. 1. The architecture of an LSTM block

1.4 L/s. Given the size of the network and to decrease the complexity of the training, the sectorization in 13 zones proposed by Rodríguez-Argote et al. (2023) (see Figure 2b) is used where nodes corresponding to the same zone are located physically close to each other and thus will have similar leakage-signatures.

For the LSTM to provide a satisfactory response it has to be trained using sufficient data of both normal (leak-free) and anomalous (leaky) conditions. Since it is non-viable to produce such amount of data experimentally, synthetic data is generated using an EPANET® simulation hydraulic model of the system (see Rossman et al. (2020)). Using this sectorization scheme, datasets $\mathbf{X} \in \mathbb{R}^{n\times m}$ containing the simulated value for n sensors at m time-steps are constructed as follows:

- For the leak-free scenario, hydraulics are simulated under nominal demands for a duration of 24 hours. Pressure head and flow rate at some selected nodes are measured at a sampling rate of 1 h.
- For the leak scenario, hydraulics are simulated under nominal demands for a duration of 12 hours. Then, from hour 13 to 24, a leak with magnitude $Q_{\text{leak}} = 1.4\,\text{L/s}$ is induced at some node. Pressure head and flow rate at some selected nodes are measured at a sampling rate of 1 h.

Two remarks are of importance regarding the construction of the “leaking” datasets: 1) Nodes with sensors and the reservoir are excluded, and 2) The reason that leaks are induced after a duration of 12 hours is that it is expected that the LSTM learns to recognize patterns in the data before and after the leaks are induced.

To provide enough realism to the synthetic data, it is artificially contaminated with uncertainty as follows. First, before the simulation begins:

- The base outflow at demand nodes are multiplied with a random uncertainty value.
- The demand multiplier for each time step of the simulation is also multiplied with a random uncertainty value.

Then, after the simulation is completed, the raw readings for each sensor are contaminated with a random uncertainty value. Uncertainty for each case is considered from ±5% to a maximum of ±25% of the nominal value. To increase the sensitivity of the LSTM to detect and learn patterns associated to leak events, residual measurements are used for training. First, a nominal dataset $\mathbf{X}_{\text{nom}} \in \mathbb{R}^{n\times m}$ is generated simulating leak-free conditions with no uncertainty. Then, the residual datasets are computed as

$$\mathbf{X}_r = \mathbf{X} - \mathbf{X}_{\text{nom}}. \tag{7}$$

The datasets are labeled into their corresponding category (zone). Datasets with leaks are categorized into 13 different zones and for datasets where no leak is found they are assigned category `'0'`. Thus, 14 different categories are proposed. Figure 3 shows an histogram presenting the target distribution of the datasets used for the training of the LSTM.

3.2 Proposed LSTM architecture

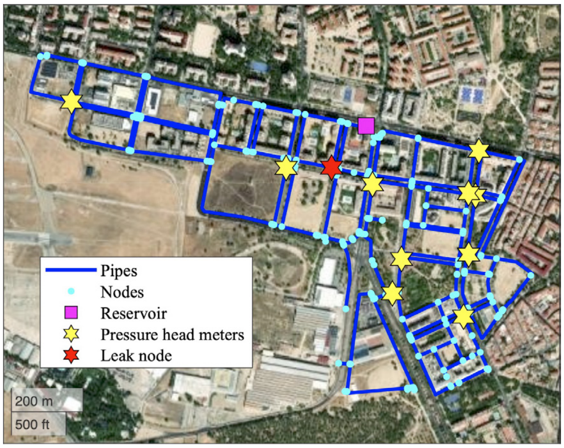
In this paper we propose the use of a LSTM neural network with categorical targets to locate leak events in a real life water distribution system given that this architecture is primarily oriented to the processing of time-dependant data such as the time series measurements obtained from pressure and flow sensors. The proposed architecture is presented in Figure 4.

The first layer is a `Sequence Input Layer` of size 11, given that 11 measurements are available. Measurements are fed into the second layer which is a `LSTM Layer` with 24 hidden units. The output of the `LSTM Layer` is fed into a `Fully Connected Layer` of size 14 (given that the inputs will be classified into 1 of 14 different categories). Then, a `Softmax Layer` is used to compute the inputs into the probability of membership to each category. Finally, a `Classification Layer` provides the highest-ranked membership as the corresponding class of the input. The proposed LSTM was implemented using the Deep Learning Toolbox® for MATLAB®. The training was performed using the hardware and software presented in Table 1.

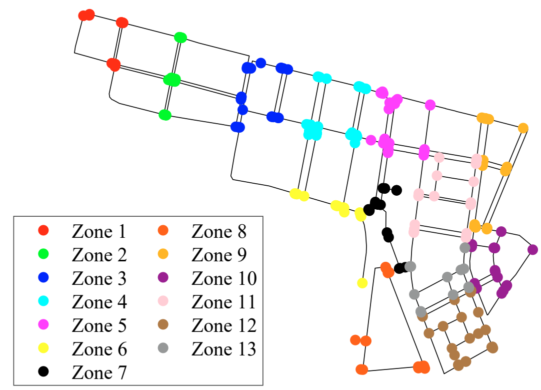
Table 1. Hardware & software used for the training of the LSTM network.

OS	Windows® 11 Pro 22H2
MATLAB® Version	R2022a
CPU	AMD Ryzen® 5700 3.8 GHz
GPU	NVIDIA® GeForce RTX 3060
RAM	64 GB DDR4 2400 MHz

The Adam optimizer was set to train the LSTM using a mini batch size of 500, initial learning rate of 0.001 and 15 000 training epochs. These hyper-parameters were selected after different trial-and-error tests given that they provide the best trade-off between training time, accuracy and test validation. Traning was performed using the GPU as execution environment. Given that in real world implementations one of the most relevant aspects of uncertainty are related to the variation in user demands, several tests will be conducted considering that *nominal* demand conditions vary between some uncertain value. Six training tests were performed: the first training test considers only datasets without uncertainty while the



(a) Geographical layout



(b) Sectorization

Fig. 2. Madrid DMA.

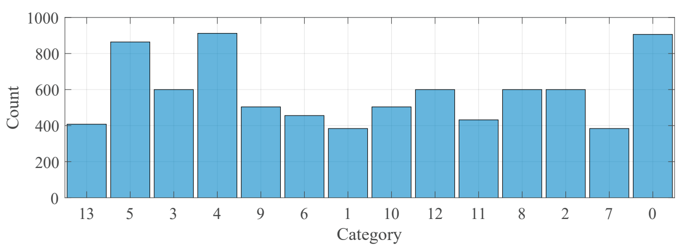


Fig. 3. Histogram of the distribution of the targets in the training sets.

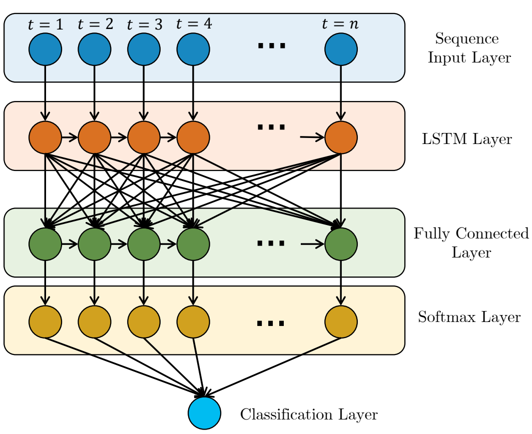


Fig. 4. Proposed architecture of the LSTM.

remaining five training tests consider datasets with a maximum uncertainty of ±5%, ±10%, ±15%, ±20% and ±25%, respectively.

4. RESULTS AND DISCUSSION

Figure 5 presents an example of the results of the training which results in a training time of 18 minutes for ±0% uncertainty and 77 minutes for ±25% uncertainty.

The accuracy of the trained LSTMs was tested by predicting the expected classes of datasets constructed specifically with this objective. Results can be consulted in Table 2, obtaining a maximum accuracy of 98.30% in correctly classifying the leak location when a new dataset is presented to the trained neural network.

Table 2. Accuracy results for the trained LSTM networks (synthetic datasets).

Uncertainty	Classification accuracy
±0%	98.30 %
±5%	96.88 %
±10%	97.54 %
±15%	96.95 %
±20%	96.93 %
±25%	97.82 %

4.1 Experimental validation

Experimental validation was performed using measurements from the operation of a Madrid DMA for a period of three days (july 17, 2019 – july 20, 2019). In this dataset, a leak of magnitude $Q_{\text{leak}} \approx 1.4\,\text{L/s}$ is induced at a node 4 at the third day of the experiment. Figure 6 shows the measurements for a) Pressure head, b) Demanded flow rate from the reservoir and c) Leaked outflow under this induced scenario. Measurements are obtained at a sampling rate of 2 minutes and thus for the period of 3 days, a total of 2160 samples are collected over a period of 4320 minutes.

Given that the LSTM was trained using datasets with measurements at a sampling rate of 1 hour, a re-sampling of the plots presented in Figure 6 was performed considering the mean value of an hour (30 samples) as an unique re-sampled piece of data. Re-sampled plots are presented in Figure 7.

Residuals of the plots presented in Figure 7 are computed as shown in Equation (7) with a dataset $\mathbf{X}_{\text{nom}}$ of dimension 11×72 under nominal conditions. Residual measurements for the experimental validation are presented in Figure 8.

Experimental validation was performed using the six trained networks. Results are summarized in Table 3.

As it can be seen, out of the six tested networks, one correctly classified the zone where the leak is occurring and two networks mistakenly classified the leak to occur in zone 5 which is located right next to the real sector and can still be considered as solutions with an acceptable margin of error given the size of the network.

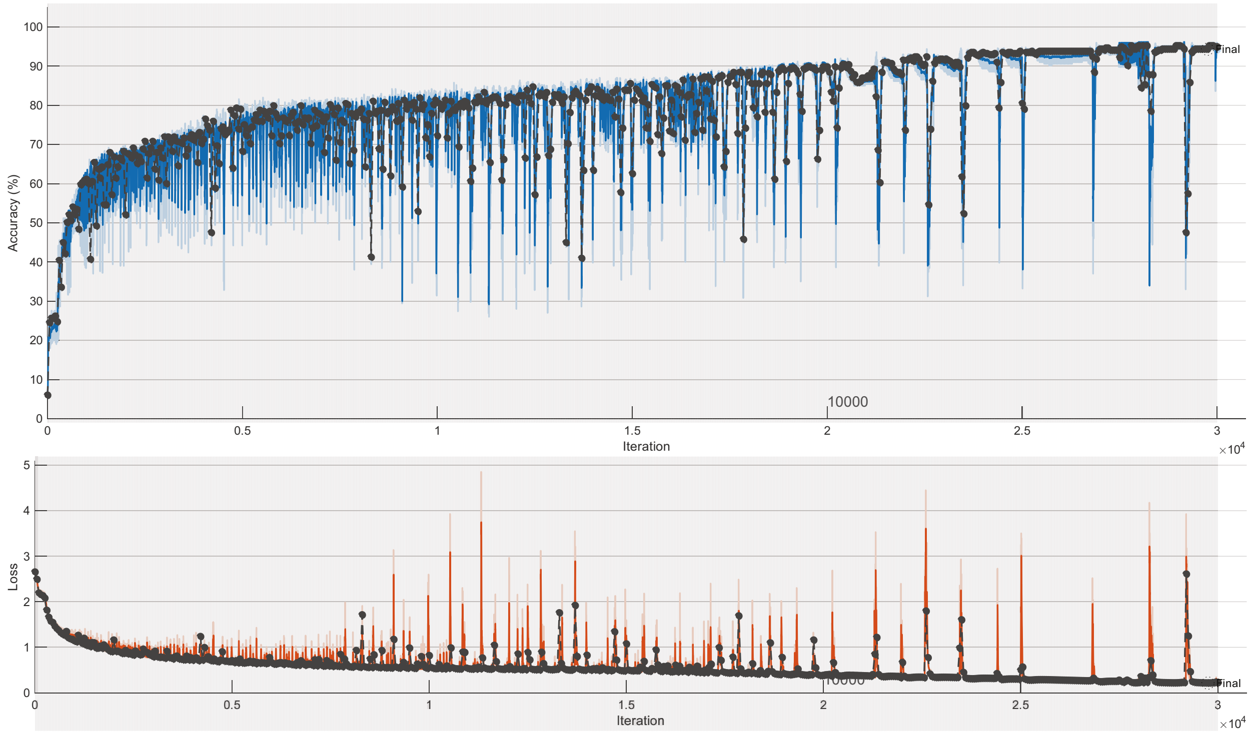


Fig. 5. Training plots of the LSTM Neural Network.

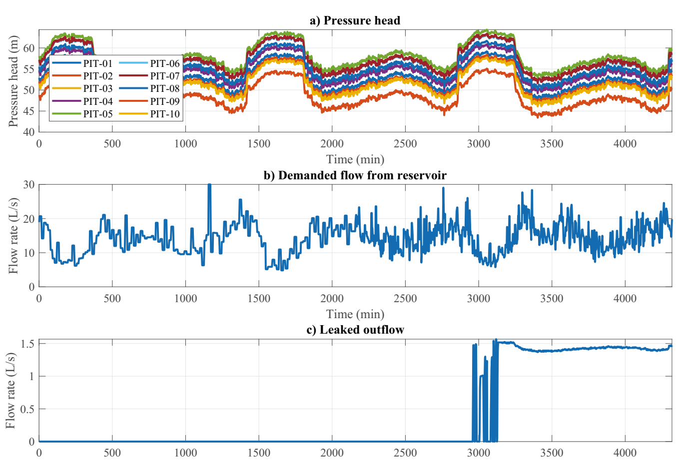


Fig. 6. Measurements of the hydraulic variables under the induced leak event.

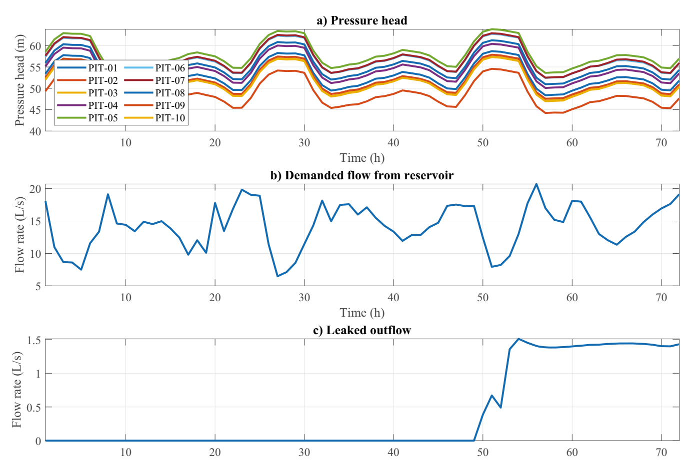


Fig. 7. Resampled measurements under the induced leak event.

Table 3. Results of the experimental validation.

LSTM Network	Real sector	Predicted sector
±0 %	4	0 (No leak)
±5 %	4	5
±10 %	4	4
±15 %	4	8
±20 %	4	9
±25 %	4	5

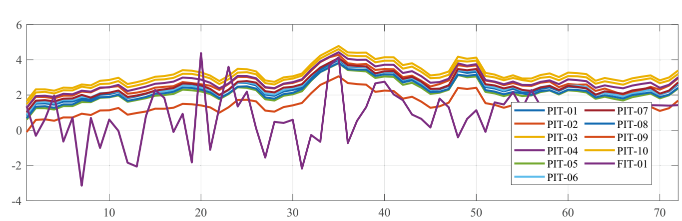


Fig. 8. Residual measurements under the induced leak event.

5. CONCLUSION

This paper presented the implementation of a LSTM Deep Neural Network for leak localization in a real WDN. The system used in this study is a DMA in Madrid, Spain. Given the size of the WDN, the number of nodes where the leak can be located and to simplify the architecture of the required LSTM, a sectorization of the WDN was proposed clustering together nodes that have similar leak signatures given their physical proximity. Given that measured data in leak scenarios for this DMA is scarce, synthetic datasets were generated using simulation software with an adequate and calibrated model of the system. For increased realism, simulations were artificially contaminated with various levels of uncertainty both in the model (user demands, consumption pattern) and in the accuracy of the sensors. Both simulation and experimental results were presented. For the case of experimental results a maximum accuracy of 98.30% and minimum accuracy of 96.88% were obtained. For the experimental validation, one of the six trained networks classified correctly the zone where the leak event is taking place, whereas two networks mistakenly classified the leak in the next-closest zone, which are still considered results with an acceptable margin of error.

Future implementations should include further analysis in how the granularity of the sectorization affects the accuracy of the LSTM in the leak localization, as well as testing more thoroughly the performance of the proposed method using real measurements. However, this would require coordination with the water-management organism for the creation of datasets considering the implementation of more leak scenarios. It is also proposed that the issue of multi-leaks should be somehow addressed, as well as testing the performance of the method using sequential targets rather than simple categorical targets with the aim of not only providing the highest ranked leak locations but also the moment at which the leaks began.

ACKNOWLEDGEMENTS

This work was developed within the framework of RICCA "Red Internacional de Control y Cómputo Aplicados". Thanks to CONAHCYT and Tecnológico Nacional de México for the funding granted for this research. We would also like to thank the project PID2020-115905RB-C21 (L-BEST) funded by MCIN/AEI/10.13039/501100011033 for supporting this research.

REFERENCES

Aggarwal, C.C. (2018). *Neural Networks and Deep Learning*. Springer International Publishing. doi:10.1007/978-3-319-94463-0.

Carnero, M.C. and Gómez, A. (2018). Optimization of maintenance in production and storage systems for domestic water. *Water Resources Management*, 32, 359–380. doi:10.1007/s11269-017-1815-2.

Fu, G., Jin, Y., Sun, S., Yuan, Z., and Butler, D. (2022). The role of deep learning in urban water management: A critical review. *Water Research*, 223, 118973. doi: https://doi.org/10.1016/j.watres.2022.118973.

Huang, S.C., Lin, W.W., Tsai, M.T., and Chen, M.H. (2007). Fiber optic in-line distributed sensor for detection and localization of the pipeline leaks. *Sensors and Actuators, A: Physical*, 135, 570–579. doi: 10.1016/j.sna.2006.10.010.

Javadiha, M., Blesa, J., Soldevila, A., and Puig, V. (2019). Leak localization in water distribution networks using deep learning. In *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, 1426–1431. doi: 10.1109/CoDIT.2019.8820627.

Kong, Y.L., Huang, Q., Wang, C., Chen, J., Chen, J., and He, D. (2018). Long short-term memory neural networks for online disturbance detection in satellite image time series. *Remote Sensing*, 10(3). doi:10.3390/rs10030452.

Martini, A., Troncossi, M., and Rivola, A. (2017). Leak detection in water-filled small-diameter polyethylene pipes by means of acoustic emission measurements. *Applied Sciences (Switzerland)*, 7. doi:10.3390/app7010002.

OECD (2016). *Water Governance in Cities*. doi: https://doi.org/https://doi.org/10.1787/9789264251090-en.

Rodríguez-Argote, C.A., Begovich-Mendoza, O., Navarro-Díaz, A., Santos-Ruiz, I., Puig, V., and Delgado-Aguiñaga, J.A. (2023). Two-leak isolation in water distribution networks based on k-nn and linear discriminant classifiers. *Water*, 15, 3090. doi: 10.3390/w15173090.

Romano, O. and Akhmouch, A. (2019). Water governance in cities: Current trends and future challenges. *Water (Switzerland)*, 11. doi:10.3390/w11030500.

Romero, L., Blesa, J., Puig, V., and Cembrano, G. (2022). Clustering-learning approach to the localization of leaks in water distribution networks. *Journal of Water Resources Planning and Management*, 148(4), 04022003. doi:10.1061/(ASCE)WR.1943-5452.0001527.

Romero, L., Blesa, J., Puig, V., Cembrano, G., and Trapiello, C. (2020). First results in leak localization in water distribution networks using graph-based clustering and deep learning. *IFAC-PapersOnLine*, 53(2), 16691–16696. doi: https://doi.org/10.1016/j.ifacol.2020.12.1104. 21st IFAC World Congress.

Rossman, L.A., Woo, H., Tryby, M., Shang, F., Janke, R., and Haxton, T. (2020). EPANET 2.2 User Manual. Technical Report EPA/600/R-20/133, U.S. Environmental Protection Agency, Washington, DC.

Salehi, M. (2022). Global water shortage and potable water safety; today's concern and tomorrow's crisis. *Environment International*, 158, 106936. doi: https://doi.org/10.1016/j.envint.2021.106936.

Sangiorgio, M., Dercole, F., and Guariso, G. (2021). Forecasting of noisy chaotic systems with deep neural networks. *Chaos, Solitons & Fractals*, 153, 111570. doi: https://doi.org/10.1016/j.chaos.2021.111570.

Santos-Ruiz, I., López-Estrada, F.R., Puig, V., Pérez-Pérez, E.J., Mina-Antonio, J.D., and Valencia-Palomo, G. (2018). Diagnosis of Fluid Leaks in Pipelines Using Dynamic PCA. volume 51, 373–380. Elsevier B.V. doi: 10.1016/j.ifacol.2018.09.604.

Santos-Ruiz, I., López-Estrada, F.R., Puig, V., and Valencia-Palomo, G. (2023). *Leak localization in water distribution networks using machine learning based on cosine features*, chapter 6, 119–145. Number SID713 in Investigación y Desarrollo. Universidad Nacional Autónoma de México. doi: 10.22201/iingen.9786073073592e.2023.

Wang, M.D., Lin, T.H., Jhan, K.C., and Wu, S.C. (2021). Abnormal event detection, identification and isolation in nuclear power plants using lstm networks. *Progress in Nuclear Energy*, 140, 103928. doi: https://doi.org/10.1016/j.pnucene.2021.103928.

Zhou, X., Tang, Z., Xu, W., Meng, F., Chu, X., Xin, K., and Fu, G. (2019). Deep learning identifies accurate burst locations in water distribution networks. *Water Research*, 166, 115058. doi: https://doi.org/10.1016/j.watres.2019.115058.